Szentkereszti Gábor

Innovations in Actuarial Mortality Forecasting

Doctoral School of Economics and Business Informatics

Supervisor: Vékás Péter, Ph.D.

Corvinus University of Budapest

Doctoral School of Economics and Business Informatics

Innovations in Actuarial Mortality Forecasting

Doctoral Dissertation

Szentkereszti Gábor

Budapest, 2026

Table of Contents

List of Figures

List of Tables

Declaration on the Use of Generative Artificial Intelligence

I declare that I used OpenAI Codex, based on GPT-5, during the preparation of this dissertation for limited language editing, proofreading, structural review, and LaTeX formatting assistance.

The generative AI system was applied only to text and substantive content originally prepared by the author. It was not used to generate research data, empirical results, mathematical proofs, or scientific conclusions. All AI-assisted output was critically reviewed, verified, and revised by the author, who assumes full responsibility for the accuracy, integrity, and final content of the dissertation.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Péter Vékás, for his continuous guidance, valuable advice, and support throughout my doctoral studies and the preparation of this dissertation. I am particularly grateful that his guidance has accompanied my academic work since my MSc thesis, for which he also served as my supervisor.

I am also grateful to the Doctoral School of Economics and Business Informatics at Corvinus University of Budapest for providing a supportive academic environment and the opportunity to pursue my doctoral research.

Chapter 1

Introduction

1.1 Mortality in historical perspective

To describe the improvement in overall mortality, Omran (1971) introduced the concept of epidemiological transition. According to him, the process of mortality improvement can be divided into three main stages. In the first stage (age of pestilence and famine), mortality is characterized by cyclicality and greater fluctuations, with major determinants being epidemics, wars, and famine. Life expectancy at birth is estimated to be between 20 and 40 years. In the next stage (age of receding pandemics), infectious diseases begin to recede, and famine ceases. Life expectancy increases to around 50 years. Population growth accelerates during this period. Meanwhile, in the final stage (age of degenerative and man-made diseases), infectious diseases are completely displaced, and degenerative diseases become the primary cause of death. Fertility becomes the critical factor for population growth.

Figure 1.1 illustrates the long-run improvement in mortality associated with these transitions through the evolution of life expectancy at birth in four European countries.

As a result of the epidemiological transition, mortality rates from infectious and respiratory diseases decline, while rates of death from degenerative and cardiovascular diseases increase. These changes in mortality patterns reflect the factors discussed above. As infant mortality declines and a larger share of the

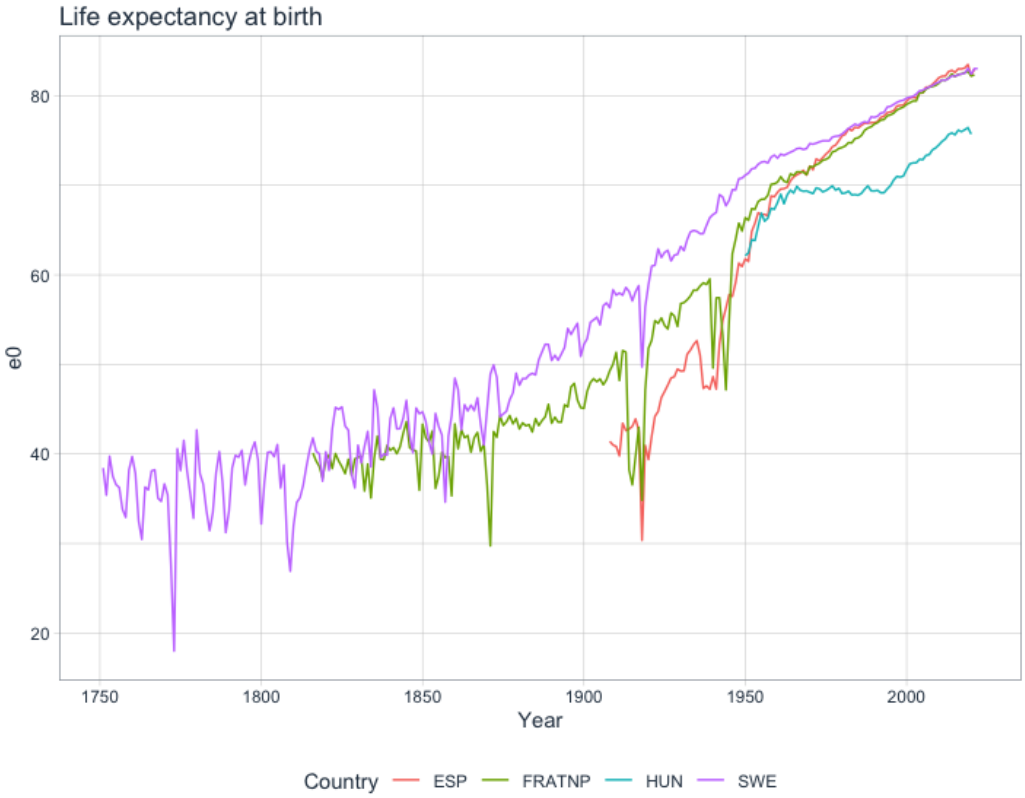

Figure 1.1: Illustration of life expectancy at birth for some countries (Source: Human Mortality Database)

population survives to older ages, cardiovascular diseases and cancers account for an increasing proportion of deaths. Omran (1971) observed this phenomenon globally, albeit with different starting points. In Northern Europe, it began at the end of the 18th century, in other parts of Europe during the 19th century, while in Asia, South America, and Africa, this process started in the second half of the 20th century. Mortality improvement during the 19th century could be attributed to improvements in hygiene and public health, as well as to mutations in pathogens. Then, in the 20th century, factors such as improved quality of life, advances in medicine (the 1970s saw the so-called cardiovascular revolution), and health-conscious lifestyles played major roles (Őri and Spéder, 2020).

Analyzing European countries, we can observe that around 1950, the northern

and western European countries had significantly better life expectancies than the southern and eastern countries. Then, over the next 20 years, several Southern and Eastern European countries experienced partial convergence. The next turn of events occurred when a crisis emerged in the states belonging to the Soviet Union, leading to a stagnation in life expectancy growth, and in some countries, even worsening mortality indicators among middle-aged male populations (Meslé and Vallin, 2002). This phenomenon lasted approximately from 1965 to 1995 in the case of the Eastern Bloc. Similar processes, to varying extents and at different times, were observed in other European countries as well. The deteriorating mortality data could be attributed to sedentary lifestyles, smoking, alcohol consumption, and poor nutrition. In addition to these factors, infectious diseases had already been minimized, and the Eastern countries were only later affected by the cardiovascular revolution, so the negative factors could not be compensated for by positive effects (Józan, 2002).

Omran's original theory was subsequently extended to account for later developments, including the cardiovascular revolution. Vallin and Meslé (2004) developed a theory consisting of three stages similar to Omran's, taking into account the processes known since then. The first stage, characterized by the decline of infectious diseases, broadly corresponds to Omran's epidemiological transition. The second stage is the cardiovascular revolution, which began around the 1970s. The final stage involves efforts to delay ageing and achieve further increases in life expectancy.

1.2 Relevance of mortality forecasting

From a macroeconomic perspective, the sustainability of pension systems has become an increasingly pressing issue in most countries. Uncertainty in mortality patterns complicates expenditure forecasting, which in turn makes the formulation of appropriate policy decisions highly challenging. In the case of pay-as-you-go pension schemes, long-term planning is indispensable. This requires the establishment of expectations regarding the best estimates of mortality and population dynamics, as only then can expected revenues and expenditures be projected for

future years. The main problem with pay-as-you-go systems is the continuous increase in life expectancy. As a result, pension funds face a growing financial burden due to prolonged benefit payments, while the persistently low fertility rate places pressure on the contribution side as well, further complicating the fulfillment of obligations.

Beyond the public pension system, the evolution of mortality rates across different birth cohorts is also highly significant for the life insurance sector. Life insurance premium calculation follows the principle of equivalence, meaning that the present value of expected revenues is set equal to the present value of expected expenditures. Since both expected revenues (premiums are paid only if the policyholder survives) and expected expenditures (the insured event is tied to survival or death) depend on mortality, accurate forecasts of future mortality rates are crucial for insurers. In most cases, premium calculations are carried out using static methods, specifically period mortality tables. This static approach assumes that mortality probabilities remain constant over time. However, due to the continuous improvement in mortality, this assumption is evidently flawed. This becomes particularly problematic for insurance products that are financially unfavorable for insurers if policyholders live longer. Such products are primarily annuities and endowment insurances. In contrast, dynamic premium calculation involves forecasting mortality probabilities for each age group and then using these projected probabilities in premium setting. This approach is especially important for insurers because life insurance contracts are typically long-term, requiring reliable assumptions for the coming decades.

In addition to premium setting, accurate mortality assumptions are also essential for the valuation of life insurance contracts, as only then can realistic projections be made about the expected profitability of portfolios. Insurers construct product-specific cash-flow models to evaluate their life insurance portfolios. These models serve to estimate expected revenues and expenditures for both current and future contracts. Naturally, as noted in relation to premium calculation, these projected cash flows are also dependent on mortality assumptions. Consequently, results can vary considerably depending on which mortality parameters are applied in the models. Overall, the forecasting of mortality rates plays a central role in cash-flow modeling and, within that, in profitability testing.

The risk arising from unexpectedly low mortality or unexpectedly high survival rates is referred to in the literature as longevity risk. Beyond state pension systems and insurance companies, longevity risk is also a pressing issue from the individual's perspective. This is because, in addition to public pensions, supplementary private pension savings—as part of self-provision—must also account for the ongoing extension of life expectancy.

The impact of longevity risk has already been recognized by regulators. Within the European Union, the determination of solvency capital requirements for insurance companies has, since January 1, 2016, been governed by the Solvency II Directive. Solvency II identifies various risks faced by insurers and requires the calculation of solvency capital needs under different scenarios. It defines risk modules and submodules, among which the longevity risk submodule of the life insurance module is particularly relevant for the present discussion. Under the regulation, longevity risk is quantified as the decrease in net asset value in the event of a permanent 20% reduction in mortality rates. From this perspective, it is reassuring that the risk is addressed in terms of corporate stability and solvency, ensuring that policyholders are not adversely affected in terms of benefits. However, it is also clear that this represents only a risk management tool. The earlier examples illustrate that insurers need to address longevity risk from many more angles.

In conclusion, it is evident that longevity risk is a significant and complex phenomenon that affects not only individuals, but also governments, private companies, and regulators. This makes it essential to employ mathematical and statistical methods in order to develop the most accurate expectations concerning the future evolution of mortality rates. These are precisely the practical challenges that our research seeks to address by providing methodological solutions.

1.3 Research questions and hypotheses

This thesis aims to address the following research questions.

- Research question 1: Does using more complex time series methods to forecast the Lee–Carter model's mortality index lead to improved forecast

accuracy? Which method provides the most reliable forecasts for actuaries and demographers?

- Hypothesis 1: Forecast accuracy can be improved by using spatiotemporal econometric models, which utilize not only the temporal information but also the spatial information contained in the data.

- Research question 2: To what extent can mortality forecasts for the Visegrád Group countries be improved by using neural networks compared to classical methods? Are the forecast results significantly better?

- Hypothesis 2: Mortality rate forecasts for Visegrád Group countries can be significantly improved using neural networks.

- Research question 3: Can neural networks produce coherent mortality forecasts? If so, under what conditions?

- Hypothesis 3: The coherence of mortality forecasts generated by a neural network depends on identifiable restrictions on its architecture and activation functions.

1.4 Structure of the thesis

Chapter 2 provides a review of the literature relevant to the methods applied in the thesis. It covers classical mortality forecasting approaches, introduces machine learning techniques, discusses their application to mortality forecasting, and presents key concepts from spatial statistics.

Chapter 3 provides a comprehensive mathematical description of the methods applied in the subsequent analysis.

Chapter 4 focuses on identifying alternative spatial and temporal time series as well as econometric methods that could be applied to forecast the mortality index of the Lee–Carter model, beyond the commonly used approaches in practice. Particular emphasis is placed on methods capable of incorporating spatial dependencies, thereby allowing for the integration of additional information into the modeling framework.

Chapter 5 presents a case study examining the performance of neural network–based mortality forecasting methods, which have become increasingly prevalent in the international literature, in the context of Visegrád Group countries. Special attention is given to Hungary, with the aim of assessing whether and under what conditions, beyond classical approaches, actuaries and demographers may benefit from experimenting with machine learning techniques when forecasting mortality rates. Moreover, I consider it essential not only to provide an aggregate assessment of model performance but also to conduct a more granular analysis, identifying for which age groups and populations specific methods yield superior results.

Chapter 6 introduces a study conducted with my co-authors, in which we demonstrate that certain neural network architectures are capable of producing coherent forecasts. Coherence is a crucial property in multipopulation models; however, this criterion has not yet been thoroughly examined in the context of machine learning approaches within the international literature.

Chapter 2

Literature review

The review is organized around the three methodological themes underlying the thesis. Section 2.1 reviews classical single- and multipopulation mortality models. Sections 2.2 and 2.3 discuss machine-learning methods and their applications to mortality forecasting. Section 2.4 reviews spatial and spatio-temporal approaches relevant to modeling dependence across populations.

2.1 Mortality forecasting

In the early periods, mortality was described using so-called mortality laws. de Moivre (1752) formulated the distribution function of mortality without parameters. Gompertz (1825) introduced a two-parameter function to model mortality, assuming that the force of mortality increases exponentially with age. Makeham (1867) extended this approach by adding a constant parameter. This additive constant accounts for age-independent factors affecting mortality.

The most widely used mortality forecasting model is associated with the work of Lee and Carter (1992). The Lee–Carter model represents the evolution of age-specific mortality through an age profile, an age-specific sensitivity term, and a common time-varying mortality index. The mortality index is then extrapolated using a time-series model to obtain forecasts of age-specific mortality rates. A formal definition of the model, including its identification constraints and estimation procedure, is provided in Section 3.1.

Lee and Carter estimated the parameters using the method of least squares. Because the Lee–Carter specification is bilinear in its unknown age- and time-specific parameters, it cannot be estimated directly as an ordinary linear regression. Under the model's identification constraints, Lee and Carter (1992) obtained a least-squares solution using singular value decomposition. From a practical perspective, they demonstrated the performance of their model using U.S. mortality data from 1933 to 1987, grouped into five-year age categories. They then presented the estimated values on a test set, as well as the projected trajectory of the time-dependent parameter up to 2065. Expected central mortality rates and life expectancies were also computed through 2065.

Wilmoth (1993) suggested an alternative estimation approach for the Lee–Carter model, replacing the SVD method with weighted least squares. He also argued that the Lee–Carter model could be applied to cause-specific mortality rates, though in such cases some causes of death may exhibit zero frequencies at certain ages, making Lee and Carter's original estimation method inapplicable. This issue is referred to as the "zero-cell" problem. Moreover, he emphasized that weighted least squares provides a better model fit at ages where the number of deaths is small. Wilmoth also presented empirical analyses based on Japanese mortality data.

Lee and Miller (2001) proposed that the adjustment of the time-dependent parameter should be based on life expectancy at birth rather than on the number of deaths. Brouhns et al. (2002) pointed out that the assumptions of normally distributed and homoscedastic error terms in the Lee–Carter model are unrealistic, and therefore suggested an alternative approach. They assumed that the number of deaths corresponding to each calendar year and age group follows a Poisson distribution, with model parameters estimated by maximum likelihood. This approach is referred to as the Poisson Lee–Carter model.

Booth et al. (2002) recommended retaining additional components after the SVD decomposition in the Lee–Carter model, thereby increasing the proportion of explained variance. In addition, they proposed alternative methods for adjusting the mortality index. Their modifications were illustrated using Australian data over different time periods, with comparisons made against the original Lee–Carter model.

In contrast to earlier approaches, Currie et al. (2004) modeled smoothed mortality rates using P-spline functions. Hyndman and Ullah (2007) also applied smoothing techniques to mortality rates. In addition, they extended the classical Lee–Carter model by incorporating multiple components into the equation and forecasted the time-dependent parameter using both ARIMA models and exponentially smoothed state-space models.

Koissi et al. (2006) compared the performance of the Lee–Carter model for Scandinavian countries using three different estimation methods: the classical SVD, maximum likelihood, and weighted least squares. They also constructed confidence intervals for the estimates using the residual bootstrap method of Efron (1979). Renshaw and Haberman (2006) extended the Lee–Carter model by incorporating cohort effects. Their empirical analysis, based on data for England and Wales from 1961 to 2003, compared annuity prices derived from different model specifications.

Cairns et al. (2006) described mortality rates using two mortality parameter vectors with a logit link function, a framework commonly referred to as the CBD model. They emphasized that their approach performs particularly well for age groups above 60. Plat (2009) further extended the CBD model by including cohort effects.

Hyndman et al. (2013) focused on the coherence of mortality forecasts. They highlighted that independently forecast mortality rates tend to diverge, whereas in the long run rates should converge. In their framework, coherence implies that age-specific mortality rates across related populations do not diverge in the long run.

Villegas et al. (2015) systematized the Lee–Carter model and its subsequent extensions, classifying them under the broader family of Generalized Age–Period–Cohort (GAPC) models.

The Lee–Carter model and the generalized age–period–cohort models are defined separately for each population. Naturally, however, there may be dependencies between mortality data across different populations. Therefore, in line with earlier approaches, researchers have developed models that account for cross-population relationships. Li and Lee (2005) introduced the Common Factor Model (CFM), which incorporates both age- and population-specific parameters as well

as common age- and period-dependent parameters. In addition, they proposed the augmented Common Factor Model (aCFM), which includes not only the common age and period parameters but also population-specific age and period parameters.

They carried out an empirical analysis forecasting mortality in 15 countries using the classical Lee–Carter model alongside their proposed CFM and aCFM models. Their results indicated that in 11 out of the 15 cases, the models incorporating common effects provided more accurate forecasts. A modification of these models was later proposed by Kleinov (2015) in the form of the Common Age Effect (CAE) model, which contains one age- and population-specific parameter as well as one age-specific and one population- and period-specific parameter.

Li and Lee (2005) forecasted mortality indices using country-specific independent autoregressive processes. Enchev et al. (2017) developed a variation of the Li–Lee model and the CAE model, experimenting with several forecasting approaches. In their work, some models are projected using multivariate random walks, while others rely on mean-reverting autoregressive processes. In multipopulation models, the common mortality index is, naturally, forecast by an independent random walk. Their analysis was carried out using mortality data from six Western European countries.

Wen et al. (2020) also constructed multipopulation models, forecasting mortality across ten different socio-economic groups in England. Their framework essentially adopts a special case of the CAE model. In addition, they outlined the interrelationships between models proposed in the literature, showing how they can be traced back to the Lee–Carter model.

According to Jarner and Jallbjørn (2020), the separately estimated LC mortality indices are cointegrated. They conducted analyses using both VAR and VECM models.

In the Hungarian literature, Májer and Kovács (2011) analyzed the impact of longevity risk using the Lee–Carter model. The authors forecasted mortality rates for older age groups and presented the net premiums of certain annuity insurance products under different forecast scenarios. Vékás (2017) applied models from the generalized age–period–cohort (GAPC) family to forecast Hungarian mortality data. Vékás (2019b) illustrated the so-called rotation phenomenon in Hungarian mortality, whereby the rate of mortality improvement slows at younger ages and

accelerates among older age groups. Vékás (2019a) provided a comprehensive overview of the GAPC model family and used it for mortality forecasting, while also presenting studies on the sustainability of the Hungarian pension system.

Arató et al. (2009) applied the Azbel model for mortality forecasting. Tóth (2021) employed the Lee–Carter and product-ratio models to analyze mortality in Central and Eastern European countries, arguing that the product-ratio model outperforms the Lee–Carter model in countries where mortality improvements were minimal during the second half of the 20th century.

2.2 Short overview of applied machine learning methods

Departing from the practical aspects of mortality forecasting, this section reviews the methodological tools applied later. A comprehensive discussion of neural networks can be found in Goodfellow et al. (2016). For the present applications, feedforward neural networks (FFNNs) and recurrent neural networks (RNNs) are of particular interest. Both types of networks belong to supervised machine learning methods and are suitable for solving classification and regression tasks. One of the earliest neural network models was introduced by Rosenblatt (1958), in which the weighted sum of inputs is passed through a sign function, thus solving a binary task.

Feedforward neural networks typically use cross-sectional data, whereas recurrent neural networks are designed for modeling sequential data. The two fundamental RNN architectures were introduced by Elman (1990) and Jordan (1986); the former feeds back the previous hidden state, while the latter feeds back the output from the previous time step. Further variants of RNNs, their training procedures, and convolutional neural networks (CNNs) are also described in Goodfellow et al. (2016).

The training of recurrent neural networks is complicated by the vanishing gradient problem, as discussed by Bengio et al. (1994) and Hochreiter and Schmidhuber (1997). In this phenomenon, long-term dependencies may be lost during training, causing relevant information to disappear. The opposite issue, known as

the exploding gradient problem, arises when large gradients make training unstable. These problems stem from the sequential nature of RNNs. Hochreiter and Schmidhuber (1997) proposed an alternative architecture called Long Short-Term Memory (LSTM), and a popular simplification of LSTM is the Gated Recurrent Unit (GRU) introduced by Cho et al. (2014). The Adam optimization algorithm, commonly used for training neural networks today, is described in Kingma and Ba (2014).

In addition to neural networks, machine learning methods have recently gained widespread popularity, even though many classical tools have been available for some time. Comprehensive discussions of machine learning methods can be found in Hastie et al. (2009) and Wüthrich and Merz (2022). Many machine learning tools are based on decision trees, which recursively split data along numeric or categorical variables to form groups that are as homogeneous as possible with respect to the target variable. Several algorithms exist for determining the splits, the most popular being ID3 and CART (Quinlan, 1986, 1996; Breiman et al., 1984).

The performance of decision trees can be enhanced using ensemble methods. In random forests, new training datasets are generated through bootstrap sampling, and variables are randomly selected at each split. Multiple decision trees are built on these datasets, and their predictions are aggregated to form the final output. Random forests are a type of ensemble method, specifically based on bagging (bootstrap aggregating) (Breiman, 1996). Both decision trees and random forests are supervised machine learning methods suitable for regression and classification tasks.

Gradient boosting methods were developed to address similar problems. The simplest form, AdaBoost, builds a sequence of shallow decision trees, with each subsequent tree focusing on correcting the errors of the previous one (Schapire, 1990; Freund and Schapire, 1996; Friedman, 2001). Unlike bagging methods, boosting constructs models sequentially, where each model depends on the previous one. Overfitting is typically controlled via a hyperparameter known as the learning rate.

Variable selection techniques applied later include LASSO and stepwise methods. LASSO is essentially a linear regression model with an added penalty term

that encourages sparsity, forcing less relevant variables to be excluded (Tibshirani, 1996). Stepwise methods also perform variable selection according to a predefined logic, most commonly through forward or backward procedures. In forward selection, one starts with an empty model and adds the variable with the greatest explanatory power iteratively until no significant variables remain. Backward selection begins with all candidate variables and removes the least significant variables one by one, refitting the model at each step. Another common approach is the best subset method, which evaluates all possible variable combinations and selects the optimal model based on a predefined information criterion (Hastie et al., 2009).

2.3 Mortality forecasting with machine learning methods

In this subsection, we provide a brief and non-exhaustive review of the related literature. Zheng et al. (2025) provide a thorough overview of deep learning methods applied to mortality forecasting. For readers seeking comprehensive introductions to actuarial applications of artificial intelligence and machine learning, including mortality forecasting, the textbooks of Wüthrich and Merz (2023) and Wüthrich et al. (2025) offer detailed theoretical and practical treatments.

2.3.1 Deep feedforward neural networks and their extensions

Beyond traditional statistical techniques, the success of machine learning—and in particular, deep learning—across multiple domains has motivated the development of more flexible modeling approaches that can automatically learn complex patterns from data in the fields of mortality forecasting and other actuarial applications (Richman, 2021).

Richman and Wüthrich (2021) proposed a deep FNN with an embedding layer to learn the non-linear mapping from the inputs of age, region, gender, and calendar year to mortality rates. Notably, their approach simultaneously fits multiple populations in one model. They found that this neural extension

achieved better out-of-sample accuracy than separate Lee–Carter models for each population.

After EIOPA (2020) and other regulators highlighted the need for transparency and caution when deploying AI models in insurance, Richman (2022) outlined several potential avenues for implementing explainable deep learning techniques so that they can be safely incorporated into the actuarial toolbox.

Scognamiglio (2022) showed how a network can be used to calibrate the Lee–Carter and Poisson Lee–Carter models for multiple populations: instead of freely predicting mortality, the network was constrained to output the Lee–Carter age and period components, thereby preserving the explainable Lee–Carter structure.

Perla and Scognamiglio (2023) proposed a framework for locally-coherent multipopulation modeling by neural networks. They assumed a Lee–Carter structure and imposed coherence within clusters of populations, performing clustering and model fitting simultaneously. They reported lower forecast errors than the Lee–Carter, Poisson Lee–Carter, and Li–Lee models while preserving an interpretable structure.

Richman and Wüthrich (2022) introduced LocalGLMnet, a neural-network-based architecture that retains the structure of generalized linear models (GLMs) while using FNNs to estimate coefficients; the model supports feature selection and enhanced interpretability. In turn, Perla et al. (2024) adapted this framework for mortality forecasting by incorporating locally connected layers that allow model coefficients to vary smoothly by age and time, producing forecasts that are both accurate and explainable.

De Mori et al. (2025) presented multi-task FNNs for multipopulation mortality forecasting. They developed several architectures with network parts that are common to all countries and thus enable the model to share information across countries as well as country-specific subparts that enable deviations from common trends, akin in spirit to the Li–Lee model. Additionally, they suggested incorporating clustering to leverage commonalities within homogeneous groups of populations.

Zhang and Zhuang (2026) applied Kolmogorov–Arnold Networks (KAN) to mortality modeling to combine predictive accuracy, smoothness control, and intrinsic interpretability. Their approach uses learnable activation functions defined

as mixtures of a Sigmoid Linear Unit (SiLU) and B-spline basis functions. They further develop a KAN-based Actuarial Neural Network (KANN) that allows initialization from classical mortality components such as Lee–Carter and Age–Period–Cohort factors. Empirical results on 34 populations from the Human Mortality Database indicate that KAN-based models achieve competitive predictive accuracy relative to established methods.

2.3.2 Recurrent neural networks

Nigri et al. (2019) applied a recurrent neural network (RNN) with a long short-term memory (LSTM) architecture that mitigates the classic RNN problem of vanishing gradients and can capture longer-term dependencies (Hochreiter and Schmidhuber, 1997). They forecasted the time-varying mortality indices of the Lee–Carter model for six countries and both genders, and achieved better accuracy than the traditional Autoregressive Integrated Moving Average (ARIMA) approach.

Bravo (2021) conducted an empirical study using Portuguese data and found that an LSTM architecture can outperform classical stochastic mortality models, producing smooth and accurate age-specific mortality forecasts across the lifespan.

Chen and Khaliq (2022) trained univariate LSTM, bidirectional LSTM, and gated recurrent unit (GRU) models on age-specific mortality series from the U.S. and compared them to the Lee–Carter benchmark in terms of out-of-sample forecasting accuracy.

As deep learning techniques require large datasets, Lindholm and Palmborg (2022) addressed the problem of data scarcity by using an efficient data utilization approach for LSTMs instead of the traditional training–validation split, combined with ensembling to improve the stability of long-term predictions.

Euthum et al. (2024) applied LSTMs and GRUs to multiple counties in Italy and compared their performance with that of three statistical benchmark models.

Among Hungarian authors, Petneházi and Gáll (2019) employed neural networks using mortality data from 40 countries. Their results indicated that LSTM networks produce more accurate forecasts than the Lee–Carter model.

2.3.3 Convolutional neural networks

Perla et al. (2021) applied convolutional neural networks (CNNs), widely used in image processing, for mortality forecasting by treating the two-dimensional grid of mortality rates by age and calendar year as a matrix to be analyzed for local patterns. Their network achieved highly accurate forecasts on data from the Human Mortality Database, and it generalized well when tested on a separate dataset of U.S. mortality.

Wang et al. (2021) applied a convolution to the age–year surface for 41 countries and demonstrated that the CNN could automatically detect cohort effects and achieved superior forecasting performance to earlier approaches.

Schnürch and Korn (2022) were the first in the literature to produce interval estimates around CNN forecasts of mortality rates. They implemented a bootstrapping-based technique for this purpose and demonstrated that it yields highly reliable prediction intervals.

2.3.4 Autoencoders

Hainaut (2018) introduced a neural network autoencoder to mortality data: in a two-stage model, an autoencoder compresses the historical age-period mortality surface into a few latent factors analogous to principal components, which are then projected forward with time-series models. This approach retained a familiar factor-decomposition interpretation while leveraging neural networks to capture non-linear features, achieving more accurate forecasts than the basic Lee–Carter on multiple countries.

Miyata and Matsuyama (2022) extended the Lee–Carter model with a variational autoencoder (VAE) as an alternative to Bayesian state space models, which enabled them to compute prediction intervals without using Markov chain Monte Carlo methods while preserving interpretability.

Tanaka and Matsuyama (2025) introduced explainable deep learning to cause-of-death mortality forecasting by proposing a convolutional autoencoder with a one-dimensional latent layer that provides a mortality index series inspired by the Lee–Carter model along with age-dependent sensitivities of cause-specific log-mortality to the time-series factor.

2.3.5 Transformers and attention mechanisms

The transformer architecture (Vaswani et al., 2017) is capable of capturing long-range dependencies in sequences and has achieved state-of-the-art results in many time-series domains. In mortality modeling, an advantage of transformers is that they can attend to relevant time points across the entire history when forecasting a particular year, rather than being constrained by sequential processing.

Roshani et al. (2022) provided one of the first applications of transformers to forecasting mortality rates by projecting the time-dependent mortality index of the Poisson Lee–Carter model. They reported that their transformer-based architecture outperformed or was at least as good as LSTMs and traditional models in terms of forecast accuracy in multiple populations.

Wang et al. (2023) implemented a transformer for mortality forecasting on several countries. Their model, which they call a time-series transformer, included a multi-head attention mechanism and positional encoding, extracting key features effectively and achieving improved performance over the Lee–Carter model and alternative neural network architectures.

2.4 Spatial statistical analysis

This section covers two topics: panel and spatio-temporal forecasting techniques, and Bayesian spatio-temporal mortality models.

2.4.1 Dynamic panel and spatio-temporal forecasting techniques

Arellano and Bond (1991) propose the Generalized Method of Moments (GMM) estimator for dynamic panel linear models, and Blundell and Bond (1998) introduce the improved System-GMM estimator. Yu et al. (2008) present the spatial dynamic panel model with fixed effects and investigate its asymptotic properties. Yu and Lee (2010) propose a spatio-temporal dynamic panel model that works for both stationary and non-stationary data. Elhorst (2014) and Salima (2018) provide broad overviews of spatial panel models, and Astuti et al. (2020) present recent theoretical developments and future research directions in this field.

Pfeifer and Deutsch (1980) introduce the Spatio-Temporal Autoregressive Moving Average (STARMA) model as an extension of the well-known ARMA framework that is also capable of capturing spatio-temporal dependencies.

Griffith (2010) and Griffith (2012) present Eigenvector Spatial Filters and Eigenvector Spatio-Temporal Filters (ESTF) for modeling spatial and spatio-temporal relationships, and present the contemporaneous and lagged spatio-temporal Moran's I as two adequate measures of spatio-temporal autocorrelation in spatial panel data. Islam et al. (2022) embed this eigenvector approach into machine learning models.

2.4.2 Bayesian spatio-temporal mortality models

We shall not consider Bayesian methods in our analysis; however, we include some examples of them here for reference. Greco and Scalone (2013), Álvaro-Meca et al. (2011), Bennett et al. (2015), Liu (2021) and Gibbs et al. (2020) apply Bayesian spatio-temporal extensions of the Lee and Carter (1992) model. Ocaña-Riola and Mayoral-Cortés (2010) estimate a hierarchical Bayesian spatial mortality model on data from the regions of Spain, while Hassan (2021) presents a spatial extension of the Lee and Carter (1992) model within a functional principal components analysis framework. Ugarte et al. (2010) propose P-spline smoothing of mortality forecasts within a Bayesian framework. Pickle (2000) embeds spatio-temporal mortality forecasting in a linear mixed effects framework. Carracedo et al. (2018) detect spatio-temporal clusters in mortality data from European countries.

Chapter 3

Applied methods for mortality forecasting

In this chapter, I summarize the most important mortality forecasting methods in the literature, with a particular focus on the models applied in the present research and case studies.

3.1 Lee–Carter model

The Lee-Carter model, introduced by Lee and Carter (1992), stands as one of the most widely adopted stochastic models for forecasting age- and time-specific mortality rates. It estimates three parameter vectors **a**, **b**, and **k**, with **a** and **b** representing age-specific parameters and **k** denoting time-specific parameters.

Let $\mathbf{M} \in \mathbb{R}^{X \times T}$ contain the age- and time-specific central mortality rates structured as follows:

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & \dots & m_{1,T} \\ \vdots & \ddots & \vdots \\ m_{X,1} & \dots & m_{X,T} \end{bmatrix}, \text{ x = 1, ..., X and t = 1, ..., T}$$

The log mortality rates estimated by Lee and Carter are given by:

$$\log(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}$$

where $\varepsilon_{x,t}$ is the residual term with zero mean and σ_e^2 variance, $\mathbf{a} \in \mathbb{R}^X$, $\mathbf{b} \in \mathbb{R}^X$ and $\mathbf{k} \in \mathbb{R}^T$.

The parameter sets (a_x, b_x, k_t) and $(a_x - cb_x, b_x, k_t + c)$ are observationally equivalent for any $c \in \mathbb{R}$. Similarly, (a_x, b_x, k_t) and $(a_x, b_x/d, dk_t)$ are observationally equivalent for any $d \in \mathbb{R} \setminus \{0\}$. Therefore, Lee and Carter proposed the following identification constraints:

$$\sum_{t=1}^{T} k_t = 0 \tag{3.1}$$

$$\sum_{x=1}^{X} b_x = 1 \tag{3.2}$$

Parameter estimation is accomplished via least squares:

$$\begin{aligned} \arg\min_{a_x, b_x, k_t} \quad & F(a_x, b_x, k_t) = \sum_{x=1}^{X} \sum_{t=1}^{T} (\log(m_{x,t}) - a_x - b_x k_t)^2 \\ \text{s.t.} \quad & \sum_{t=1}^{T} k_t = 0 \\ & \sum_{x=1}^{X} b_x = 1 \end{aligned}$$

Due to equation (3.1), $\widehat{a_x}$ simply represents the average of the log mortality rates over time.

$$\widehat{a_x} = \frac{\sum_{t=1}^{T} ln(m_{x,t})}{T}$$

Upon estimating a_x, focus shifts to b_x and k_t. Let $\mathbf{Z} \in \mathbb{R}^{X \times T}$ denote the matrix of the central logarithmic mortality rates:

$$\mathbf{Z} = \begin{bmatrix} \log(m_{1,1}) - \widehat{a_1} & \dots & \log(m_{1,T}) - \widehat{a_1} \\ \vdots & \ddots & \vdots \\ \log(m_{X,1}) - \widehat{a_X} & \dots & \log(m_{X,T}) - \widehat{a_X} \end{bmatrix}$$

Let

$$\mathbf{b} = (b_1, \dots, b_X)^\top \in \mathbb{R}^X \quad \text{and} \quad \mathbf{k} = (k_1, \dots, k_T)^\top \in \mathbb{R}^T.$$

The rank-one matrix corresponding to the age and period effects is

$$\mathbf{C} = \mathbf{b}\mathbf{k}^\top \in \mathbb{R}^{X \times T},$$

whose (x, t)th element is $C_{x,t} = b_x k_t$.

This leads to the optimization problem:

$$\arg\min_{b_x, k_t} F(\widehat{a_x}, b_x, k_t) = \sum_{x=1}^{X} \sum_{t=1}^{T} (\log(m_{x,t}) - \widehat{a_x} - b_x k_t)^2 = \|\mathbf{Z} - \mathbf{C}\|_F^2$$

where $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2}$ is the Frobenius norm.

This optimization problem can be addressed using the Singular Value Decomposition Theorem for Matrices and the Eckart-Young-Mirsky theorem (Friedberg et al., 2002). Let $\mathbf{u}_1$ be the first left singular vector, $\mathbf{v}_1$ the first right singular vector, and σ_1 the first singular value.

According to the Eckart–Young–Mirsky theorem, the best rank-one approxi-

mation of $\mathbf{Z}$ is

$$\mathbf{Z}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top.$$

This yields the estimates of $\widehat{b}_x$ and $\widehat{k}_t$, where γ is a non-zero constant:

$$\widehat{b}_x = \frac{u_{1,x}}{\gamma}, \qquad \widehat{k}_t = \gamma \sigma_1 v_{1,t},$$

or, equivalently, in vector form,

$$\widehat{\mathbf{b}} = \frac{1}{\gamma}\mathbf{u}_1, \qquad \widehat{\mathbf{k}} = \gamma \sigma_1 \mathbf{v}_1.$$

To satisfy the identification constraint $\sum_{x=1}^{X} \widehat{b}_x = 1$, we set

$$\gamma = \sum_{x=1}^{X} u_{1,x}.$$

Since the rows of $\mathbf{Z}$ are centered over time, $\sum_{t=1}^{T} v_{1,t} = 0$, and therefore $\sum_{t=1}^{T} \widehat{k}_t = 0$ also holds.

Finally, Lee and Carter proposed an adjustment for $\widehat{k}_t$. For each year t, the adjusted mortality index $\widehat{k}_t^{\text{adj}}$ is defined as the solution of

$$D_t = \sum_{x=1}^{X} D_{x,t} = \sum_{x=1}^{X} E_{x,t} \exp\left(\widehat{a}_x + \widehat{b}_x \widehat{k}_t^{\text{adj}}\right),$$

where $D_{x,t}$ denotes the observed death counts and $E_{x,t}$ denotes the corresponding exposure at age x in year t. The equation is solved numerically for $\widehat{k}_t^{\text{adj}}$.

After the adjustment, the model reproduces the observed total number of deaths in each year. According to the authors' opinion, this is necessary because the estimation method considers all age groups with equal weight, thus providing a more accurate estimation for the younger population, but the elderly contribute significantly more weight to the number of deaths.

The meaning of the parameters:

- a_x: Age-specific parameter representing the average logarithmic mortality,

i.e., the average of the logarithmic mortality rates across periods at each age.

- k_t: Mortality index, capturing the temporal variation in mortality. It depends on calendar time.
- b_x: Age-specific sensitivity to k_t, indicating how changes in k_t affect the mortality rate at age x.
- $\varepsilon_{x,t} \sim \mathcal{N}(0, \sigma^2)$: Random error term.

For forecasting, $\widehat{k}_t$ is typically modeled as a random walk with drift (RWD):

$$\widehat{k}_t = \widehat{k}_{t-1} + c + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\text{RWD}}),$$

where $c \in \mathbb{R}$ denotes the drift (trend) parameter and ε_t is a normally distributed random error term.

Then the forecast for j periods ahead is straightforward:

$$\mathbb{E}[\widehat{k}_{T+j}] = \widehat{k}_T + j\,\widehat{c}.$$

3.2 Poisson Lee–Carter model

Beyond its classical formulation, the LC model has several extensions. Brouhns et al. (2002) proposed the so-called Poisson LC model, in which the LC parameters are estimated using the maximum likelihood method. This approach automatically accounts for the weighting by population size in each age group and appropriately handles the heteroscedasticity of the error terms. In what follows, I present this model variant.

Let $D_{x,t}$ denote the number of deaths and $E_{x,t}$ the corresponding central exposure at age x in year t. Assume that the random variable $D_{x,t}$ follows a Poisson distribution with parameter $E_{x,t}\mu_{x,t}$, so $D_{x,t} \sim Pois(E_{x,t}\mu_{x,t})$ with $\mu_{x,t} = exp\{a_x + b_x k_t\}$.

The log-likelihood function of the Poisson model is:

$$\ell(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{k}) = \sum_{x,t} [D_{x,t}(a_x + b_x k_t) - E_{x,t} e^{a_x + b_x k_t}] + constant$$

where $constant = \sum_{x,t} [D_{x,t} ln(E_{x,t}) - ln(D_{x,t}!)]$

To maximize the log-likelihood function, an iterative method is necessary due to the lack of a closed-form solution. Goodman (1979) proposed an updating procedure for finding the optimal parameter vectors of these problems (the Newton-Raphson method).

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mathbf{H}_{\ell}^{-1} \left(\boldsymbol{\theta}^{(i)} \right) \nabla \ell \left(\boldsymbol{\theta}^{(i)} \right),$$

where $\mathbf{H}_{\ell}(\boldsymbol{\theta}^{(i)})$ denotes the Hessian matrix of the log-likelihood function.

At iteration i, let

$$\widehat{D}_{x,t}^{(i)} = E_{x,t} \exp \left(a_x^{(i)} + b_x^{(i)} k_t^{(i)} \right)$$

denote the fitted number of deaths at age x in year t.

After applying the procedure subject to equation (3.1) and equation (3.2), we obtain the following formulas:

$$\widehat{a}_x^{(i+1)} = \widehat{a}_x^{(i)} - \frac{\sum_t (D_{x,t} - \widehat{D}_{x,t}^{(i)})}{-\sum_t \widehat{D}_{x,t}^{(i)}}$$

$$\widehat{k}_t^{(i+2)} = \widehat{k}_t^{(i+1)} - \frac{\sum_x (D_{x,t} - \widehat{D}_{x,t}^{(i+1)}) \widehat{b}_x^{(i+1)}}{-\sum_x \widehat{D}_{x,t}^{(i)} (\widehat{b}_x^{(i+1)})^2}$$

$$\widehat{b}_x^{(i+3)} = \widehat{b}_x^{(i+2)} - \frac{\sum_t (D_{x,t} - \widehat{D}_{x,t}^{(i+2)}) \widehat{k}_t^{(i+2)}}{-\sum_t \widehat{D}_{x,t}^{(i+2)} (\widehat{k}_t^{(i+2)})^2}$$

After updating $\widehat{k}_t$, recalculate it with the formula:

$$\widehat{k}_t^{(i+1)} = (\widehat{k}_t^{(i+1)} - \frac{\sum_{t=1}^{T} \widehat{k}_t^{(i+1)}}{T}) \sum_{x=1}^{X} \widehat{b}_x^{(i)}$$

The same normalization is applied to b_x:

$$\widehat{b}_x^{(i+1)} = \frac{\widehat{b}_x^{(i+1)}}{\sum_{x=1}^{X} \widehat{b}_x^{(i+1)}}$$

With these modifications, equation (3.1) and equation (3.2) are satisfied. The value of the log-likelihood function and its increment after one iteration are:

$$\ell^{(i)} = \sum_{x,t} [D_{x,t}(\widehat{a}_x^{(i)} + \widehat{b}_x^{(i)}\widehat{k}_t^{(i)}) - E_{x,t} \exp(\widehat{a}_x^{(i)} + \widehat{b}_x^{(i)}\widehat{k}_t^{(i)})]$$

$$\Delta_\ell^{(i)} = |\ell^{(i)} - \ell^{(i-1)}|$$

Iterate the parameters until $\Delta_\ell^{(i)} < 10^{-6}$.

3.3 Li–Lee model

When Lee–Carter models are fitted independently to multiple populations, the ratios of mortality rates at the same ages but across different populations typically converge over time to zero or infinity, which is not supported by empirical evidence. To address this issue, Li and Lee (2005) proposed a coherent multipopulation extension of the LC model, in which a common long-term trend is captured by shared parameters, while short-term deviations around this trend are modeled using population-specific parameters. The Li–Lee (LL) model can be expressed in the following form. The notation follows that of the previous section:

$$\log(m_{i,x,t}) = a_{i,x} + B_x K_t + b_{i,x} k_{i,t} + \varepsilon_{i,x,t},$$

where B_x and K_t are common across populations, while $a_{i,x}$, $b_{i,x}$, and $k_{i,t}$ are population-specific parameters. The index i refers to the i-th population.

The model is estimated in two steps. First, the common factor model is estimated by solving

$$\min_{a_{i,x},B_x,K_t} \sum_i \sum_x \sum_t \left[\log(m_{i,x,t}) - a_{i,x} - B_x K_t\right]^2 .$$

Singular value decomposition (SVD) is then applied to the population-specific residuals

$$e_{i,x,t} = \log(m_{i,x,t}) - a_{i,x} - B_x K_t$$

to estimate the population-specific factors $b_{i,x}$ and $k_{i,t}$.

The common mortality index K_t is forecast using an ARIMA model, whereas the population-specific mortality indices $k_{i,t}$ are modeled using either a random walk (RW) or a first-order autoregressive process (AR(1)).

Specifically:

- Random walk (RW):

$$k_{i,t} = k_{i,t-1} + \phi_{i,t}, \qquad \phi_{i,t} \sim \mathcal{N}(0, \sigma_i^2).$$

- AR(1) process:

$$k_{i,t} = c_{i,0} + c_{i,1} k_{i,t-1} + \phi_{i,t}, \qquad \phi_{i,t} \sim \mathcal{N}(0, \sigma_i^2), \qquad |c_{i,1}| < 1.$$

Under a random walk without drift, the conditional point forecast remains constant at its last observed value. Under an AR(1) process with an autoregressive coefficient satisfying $|c_{i,1}| < 1$, the point forecast converges to the population-specific long-run mean

$$\frac{c_{i,0}}{1 - c_{i,1}}.$$

Consequently, the population-specific components do not generate diverging point forecasts in the long run.

The formal definition of coherence and its implications for stochastic mortality forecasts are discussed in Section 6.3.1.

3.4 Recurrent neural networks

In the following subsection, I briefly present the three recurrent neural networks that I apply in the later chapters.

3.4.1 Simple recurrent neural networks

Rumelhart et al. (1986) introduced a simple recurrent neural network (RNN) structure suitable for modeling sequential data, along with the backpropagation-through-time method for training it.

Let the input vector at a given time step be $x^{(t)}$, the previous hidden state of the network $h^{(t-1)}$, and the output $y^{(t)}$. Let W_x, W_h, and W_y denote the weight matrices associated with these components of the model, b_x and b_y the additive constants for the input and output, and f_h and f_y the activation functions. Using these notations, the RNN structure can be described as follows:

$$h^{(t)} = f_h\big(W_h h^{(t-1)} + W_x x^{(t)} + b_x\big),$$

$$y^{(t)} = f_y\big(W_y h^{(t)} + b_y\big).$$

Due to the so-called vanishing and exploding gradient phenomena, training these networks in practice is difficult, and Bengio et al. (1994) pointed out that classical RNNs cannot capture long-term dependencies. In the following, we describe two RNN structures that are more widely used to address these problems.

3.4.2 Long short-term memory

Hochreiter and Schmidhuber (1997) proposed an alternative architecture that, thanks to its 'gated' design, is able to store useful long-term information. This

is called a Long Short-Term Memory (LSTM) network. Retaining the previous notations, the LSTM model can be described by the following equations:

Forget gate:

$$f^{(t)} = \sigma\big(W_{f,h}h^{(t-1)} + W_{f,x}x^{(t)} + b_f\big),$$

Input gate:

$$i^{(t)} = \sigma\big(W_{i,h}h^{(t-1)} + W_{i,x}x^{(t)} + b_i\big),$$

Output gate:

$$o^{(t)} = \sigma\big(W_{o,h}h^{(t-1)} + W_{o,x}x^{(t)} + b_o\big),$$

Candidate cell state:

$$\tilde{c}^{(t)} = \tanh\big(W_{c,h}h^{(t-1)} + W_{c,x}x^{(t)} + b_c\big),$$

Cell state update:

$$c^{(t)} = f^{(t)} * c^{(t-1)} + i^{(t)} * \tilde{c}^{(t)},$$

Hidden state:

$$h^{(t)} = o^{(t)} * \tanh(c^{(t)}),$$

where $*$ denotes element-wise (Hadamard) multiplication, σ is the sigmoid function, and tanh is the hyperbolic tangent function.

3.4.3 Gated recurrent unit

Cho et al. (2014) introduced a gated recurrent network similar to an LSTM to capture long-term dependencies, called the Gated Recurrent Unit (GRU). The equations of the GRU network are as follows:

Update gate:

$$z^{(t)} = \sigma\big(W_{z,h}h^{(t-1)} + W_{z,x}x^{(t)} + b_z\big),$$

Reset gate:

$$r^{(t)} = \sigma\big(W_{r,h}h^{(t-1)} + W_{r,x}x^{(t)} + b_r\big),$$

Candidate hidden state:

$$\tilde{c}^{(t)} = \tanh\left(W_{c,h}(h^{(t-1)} * r^{(t)}) + W_{c,x}x^{(t)} + b_c\right),$$

Hidden state update:

$$h^{(t)} = (1 - z^{(t)}) * h^{(t-1)} + z^{(t)} * \tilde{c}^{(t)},$$

where σ is the sigmoid function, and tanh is the hyperbolic tangent function.

Chapter 4

Mortality forecasting in geographical time and space

This chapter is based on the author's previously published work Szentkereszti and Vékás (2026).

4.1 Introduction

The vast majority of researchers, actuaries and demographers use standard time series analysis techniques to project time-varying parameters of popular mortality forecasting methods such as the Lee–Carter (LC) and Li–Lee (LL) models. However, spatial dependence can be as significant as temporal autocorrelation in these time series, and the underlying panel structure of the data is often neglected.

We draw on techniques from panel and spatial econometrics, including ordinary and spatial dynamic panel linear models, spatio-temporal autoregressive integrated moving average processes, and spatial eigenvector filters, to capture such dependence and improve projections. We present a methodology to estimate the parameters of these techniques from spatial multipopulation mortality series, select their optimal hyperparameters, and use them for forecasting. We propose a tailor-made robust selection framework to identify the best model–technique combinations for each country, as well as a bootstrap-based procedure to quantify projection uncertainty with accurate nominal coverage on a separate validation

period and a strategy for assessing the quality of the resulting prediction intervals.

We test these methods on mortality data from 22 European countries. The results show that the proposed techniques yield a clear advantage in both point and interval forecasts for several populations, and these findings are corroborated by a robust selection design and additional robustness checks. These improvements have the potential to deliver meaningful gains for life insurance, pensions, and other contexts involving longevity risk.

4.2 Spatial and spatio-temporal autocorrelation

Spatial autocorrelation refers to the correlation of a variable with itself across space—a manifestation of Tobler's First Law of Geography (Tobler, 1970). Moran's I, introduced by Moran (1950), is a foundational statistic used to quantify global spatial autocorrelation, measuring the extent to which geographically proximate observations tend to exhibit similar (positive autocorrelation), dissimilar (negative autocorrelation) or spatially random (zero autocorrelation) values. The methodological foundations and theoretical properties of Moran's I were significantly advanced by Cliff and Ord (1973), establishing it as a core tool for exploratory spatial data analysis.

Given a set of N spatial observations $\{u_i\}_{i=1}^{N}$ and an $N \times N$ spatial weights matrix $\mathbf{W}_{\text{space}} \triangleq [w_{ij}]$, Moran's I is defined as:

$$I \triangleq \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(u_i - \bar{u})(u_j - \bar{u})}{\sum_i (u_i - \bar{u})^2},$$

where $\bar{u}$ is the mean of the observations, and w_{ij} encodes spatial relationships (e.g., *contiguity weights*, where $w_{ij} = 1$ if regions i and j are neighbors and 0 otherwise, when values at neighboring locations are assumed to be autocorrelated; or *inverse-distance weights*, where $w_{ij} = 1/\delta_{ij}$, when autocorrelation is assumed to decay with geographical distance δ_{ij} measured between pairs of capital cities or centroids).[1]

[1] Moran's I values are slopes and not correlation coefficients, and are thus not constrained to the interval between –1 and 1.

Local Indicators of Spatial Association (LISA, Anselin, 1995) provide a standard framework for detecting spatial clustering and spatial outliers that may be obscured in global analyses. By decomposing the global Moran's I statistic into contributions for each geographic unit, LISA identifies areas where neighboring values are significantly similar (high–high or low–low clusters) or dissimilar (high–low or low–high outliers). Statistical significance is typically assessed through permutation testing.

Spatio-temporal autocorrelation refers to the correlation of a variable with itself across space and time; in other words, it describes the similarity or dissimilarity of values measured at proximate locations and time points. In extending Moran's I to spatio-temporal settings, the spatial weights matrix $\mathbf{W}_{\text{space}}$ is expanded using Kronecker products to form a spatio-temporal weights matrix $\mathbf{W}_{\text{space-time}}$.

A common contemporaneous specification is given by:

$$\mathbf{W}_{\text{space-time}}^{(\text{contemporaneous})} \triangleq \mathbf{I}_T \otimes \mathbf{W}_{\text{space}} + \mathbf{W}_{\text{time}} \otimes \mathbf{I}_N,$$

where $\mathbf{I}_T$ and $\mathbf{I}_N$ are identity matrices of size T and N, respectively, and $\mathbf{W}_{\text{time}}$ is a temporal adjacency matrix of size T connecting each time period with its predecessor and successor, with ones on the first lower and upper off-diagonals and zeros elsewhere (Griffith, 2010, 2012; Griffith and Paelinck, 2018).

Alternatively, the lagged specification is given by:

$$\mathbf{W}_{\text{space-time}}^{(\text{lagged})} \triangleq \mathbf{W}_{\text{time}} \otimes \mathbf{W}_{\text{space}} + \mathbf{W}_{\text{time}} \otimes \mathbf{I}_N,$$

capturing the relationship between observations at each location and their neighbors' past values.

Given these weights matrices, spatio-temporal Moran's I is defined as:

$$I_{\text{space-time}}^{(\text{spec})} \triangleq \frac{NT}{\sum_{(i,t)} \sum_{(j,s)} w_{(i,t),(j,s)}^{(spec)}} \frac{\sum_{(i,t)} \sum_{(j,s)} w_{(i,t),(j,s)}^{(spec)} (u_{(i,t)} - \bar{u})(u_{(j,s)} - \bar{u})}{\sum_{(i,t)} (u_{(i,t)} - \bar{u})^2},$$

where spec $\in$ {contemporaneous,lagged} is the chosen specification, $u_{(i,t)}$ denotes observations at location i and time t, and $\bar{u}$ is the overall mean across all spatial and temporal observations (Griffith, 2010).

It is both theoretically plausible and empirically supported to assume that mortality improvements exhibit spatio-temporal autocorrelation, thereby motivating the integration of spatio-temporal econometric methods into mortality forecasting.

4.3 Forecasting techniques

We consider the following forecasting techniques for the time series of the mortality indices k_{it} of the LC or LL models, with N and T denoting the number of countries and calendar years, respectively, and the notations $\boldsymbol{\xi}_t \triangleq (\xi_{1t}, \xi_{2t}, \ldots, \xi_{Nt})^\intercal$ $(t = 1, 2, \ldots, T)$ for the time-dependent and $\boldsymbol{\xi} \triangleq (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T)^\intercal$ for the overall vector of innovations of the time series models with mean $\boldsymbol{\mu} \triangleq \mathbb{E}(\boldsymbol{\xi})$ and covariance matrix $\boldsymbol{\Sigma} \triangleq \mathbb{E}\left[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^\intercal\right]$:

1. Random Walk with Drift (RWD):

$$\begin{aligned} k_{it} &= \alpha_i + k_{i,t-1} + \xi_{it}, \\ \boldsymbol{\xi} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \boldsymbol{\mu} &= \mathbf{0}, \quad \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2 \mathbf{I}_{T-1}, \sigma_2^2 \mathbf{I}_{T-1}, \ldots, \sigma_N^2 \mathbf{I}_{T-1}), \end{aligned}$$

   so that innovations are multivariate normal with mean zero, constant variance within countries, and no correlations across both countries and time periods. Assuming that the first-order differenced series is stationary, parameters can be estimated by maximum likelihood estimation (MLE). Lee and Carter (1992) used this specification for US data, and it has been shown to perform well across a large number of populations (Tuljapurkar et al., 2000).

2. Autoregressive Integrated Moving Average (ARIMA):

$$\begin{aligned} y_{it} \triangleq \Delta^{(d_i)} k_{it} &= \alpha_i + \sum_{\ell=1}^{p_i} \phi_{i\ell} y_{i,t-\ell} + \sum_{\ell=1}^{q_i} \theta_{i\ell} \xi_{i,t-\ell} + \xi_{it}, \\ \boldsymbol{\xi} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \boldsymbol{\mu} &= \mathbf{0}, \quad \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2 \mathbf{I}_{T-d_1}, \sigma_2^2 \mathbf{I}_{T-d_2}, \ldots, \sigma_N^2 \mathbf{I}_{T-d_N}). \end{aligned}$$

Again, it is assumed that innovations are multivariate normal with mean zero, homoskedastic within countries, and uncorrelated across countries and time periods. Differencing is applied to ensure the removal of unit roots for stationarity. As with RWD, the standard estimation method is MLE. The numbers of lags p_i and q_i by country can be optimized by minimizing the Bayesian information criterion (BIC) or related alternatives. The RWD is a special case with $p_i = 0$, $d_i = 1$ and $q_i = 0$ $(i = 1, 2, \ldots, N)$. There are no interactions among cross-sectional units.

3. Dynamic Panel Linear Model (DPLM):

$$\mathbf{y}_t \triangleq \Delta^{(d)}\mathbf{k}_t = \boldsymbol{\alpha} + \sum_{\ell=1}^{p} \phi_\ell \mathbf{y}_{t-\ell} + \boldsymbol{\xi}_t,$$
$$\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2\mathbf{I}_{T-d}, \sigma_2^2\mathbf{I}_{T-d}, \ldots, \sigma_N^2\mathbf{I}_{T-d}\right).$$

It is assumed that innovations have mean zero and are homoskedastic within countries and uncorrelated across countries and time periods. This in turn amounts to uniform ARIMA specifications with $d_i = d$, $p_i = p$, $\phi_{i\ell} = \phi_\ell$, and $q_i = 0$ $(i = 1, 2, \ldots, N)$, but without the assumption of multivariate normality. Popular estimators for this model are the Generalized Method of Moments (GMM, Arellano and Bond, 1991) and System-GMM (Blundell and Bond, 1998). The lag order p can be determined based on the statistical significance of the autoregressive terms. There are no interactions among cross-sectional units.

4. Vector Autoregression (VAR):

$$\mathbf{y}_t \triangleq \Delta^{(d)}\mathbf{k}_t = \boldsymbol{\alpha} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \mathbf{y}_{t-\ell} + \boldsymbol{\xi}_t,$$
$$\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2\mathbf{I}_{T-d}, \sigma_2^2\mathbf{I}_{T-d}, \ldots, \sigma_N^2\mathbf{I}_{T-d}\right).$$

Innovations are assumed to have mean zero and equal variance within countries, with no correlation across countries and time periods. Differencing ensures the removal of unit roots, as stationarity is required to fit the model.

Estimation is typically performed by ordinary least squares (OLS), which does not require normality of the innovations. However, in our application, the ratio of observations to parameters is well below the commonly applied rule of thumb of at least five (Van Belle, 2008), so OLS is not recommended. Instead, coefficients can be estimated using elastic net — or, as in our special case, LASSO or L_1 — regularization (Guibert et al., 2019) in order to achieve a parsimonious description of the dependence structure. The lag order p can be determined by minimizing the cross-validated mean squared error (CV-MSE). The VAR framework allows for cross-sectional interactions, which are not necessarily spatial in nature.

5. Spatio-Temporal ARIMA (STARIMA, Pfeifer and Deutsch, 1980):

$$
\begin{aligned}
z_{it} &\triangleq \frac{\Delta^{(d)} k_{it} - \underset{i,t}{\text{mean}}\left(\Delta^{(d)} k_{it}\right)}{\underset{i,t}{\text{sd}}\left(\Delta^{(d)} k_{it}\right)} = \\
&= \alpha + \sum_{\ell=1}^{p} \sum_{r=0}^{\lambda_\ell} \phi_{\ell r} W^{(r)} z_{i,t-\ell} + \sum_{\ell=1}^{q} \sum_{r=0}^{\mu_\ell} \theta_{\ell r} W^{(r)} \xi_{i,t-\ell} + \xi_{it}, \\
\boldsymbol{\mu} &= \mathbf{0}, \quad \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{N(T-d)},
\end{aligned}
$$

where $W^{(r)}$ is the r-th order spatial lag operator representing adjacency of r-th order.[2] Innovations are assumed to have mean 0 and equal variance *across all countries*, with no correlation across countries or time periods. Differencing ensures the removal of unit roots. Estimation is performed using the Kalman filter (KF), and the numbers of time lags p and q and space-time lags λ_ℓ ($\ell = 1, 2, \ldots, p$) and μ_ℓ ($\ell = 1, 2, \ldots, q$) can be determined by minimizing the BIC. STARIMA is capable of capturing complex spatio-temporal interactions.

6. Spatial Dynamic Panel Linear Model (SDPLM):

$$
\begin{aligned}
&\mathbf{y}_t \triangleq \Delta^{(d)} \mathbf{k}_t = \boldsymbol{\alpha} + \phi_{\text{space}} \mathbf{W}_{\text{space}} \mathbf{y}_t + \phi_{\text{time}} \mathbf{y}_{t-1} + \phi_{\text{space-time}} \mathbf{W}_{\text{space}} \mathbf{y}_{t-1} + \boldsymbol{\xi}_t, \\
&\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\Sigma} = \text{diag}\left(\sigma_1^2 \mathbf{I}_{T-d}, \sigma_2^2 \mathbf{I}_{T-d}, \ldots, \sigma_N^2 \mathbf{I}_{T-d}\right).
\end{aligned}
$$

[2] E.g., $W^{(2)}$ takes the average value of neighbors of neighbors, excluding the unit itself.

It is assumed that innovations have mean zero and are homoskedastic within countries and uncorrelated across countries and time periods. There is no normality assumption, and parameters are estimated by quasi maximum likelihood (QML), with the Lee–Yu transformation removing finite-sample bias (Yu and Lee, 2010). The presence or absence of a time lag and a space-time lag, the choice between adjacency and inverse-distance spatial weights, and the necessity of the Lee–Yu transformation can be determined by minimizing the quasi-BIC. SDPLM allows for structured and easily interpretable spatio-temporal interactions.

7. Eigenvector Spatio-Temporal Filter (ESTF):

$$\begin{aligned}\Delta^{(d)}\mathbf{k}_t &= \boldsymbol{\alpha} + \mathbf{S}_{\text{space-time}}\,\boldsymbol{\beta} + \mathbf{S}_{\text{space}}\,\boldsymbol{\gamma} + \mathbf{P}\boldsymbol{\phi} + \boldsymbol{\xi}_t,\\ \boldsymbol{\mu} &= \mathbf{0}, \quad \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2\mathbf{I}_{T-d}, \sigma_2^2\mathbf{I}_{T-d}, \ldots, \sigma_N^2\mathbf{I}_{T-d}),\end{aligned}$$

where $\mathbf{S}_{\text{space}}$ and $\mathbf{S}_{\text{space-time}}$ are matrices of eigenvectors of the spatial and spatio-temporal weights matrices $\mathbf{W}_{\text{space}}$ and $\mathbf{W}_{\text{space-time}}$, and $\mathbf{P}$ is a matrix of country dummy variables. Griffith (2010) proposes keeping all country dummy variables in the equation and using stepwise OLS (SW-OLS) estimation for the sake of parsimony, or alternatively, shrinkage by LASSO can be applied[3]. Choices between adjacency and inverse-distance spatial weights and between contemporaneous and lagged specifications can be made by minimizing the BIC in the case of SW-OLS and the deviance ratio (DR) in the case of LASSO. ESTF allows for complex spatio-temporal interactions, but is typically less interpretable than an SDPLM.

Chief properties of the proposed techniques are displayed in Table 4.1. Additionally, we consider two ensemble techniques. The first one, denoted by ENS1, computes the mean predicted log-mortality rate across the seven forecasting techniques for every country, age, and calendar year in both the LC and LL models, whereas the second one, denoted by ENS2, computes a trimmed mean excluding the lowest and highest predictions.

[3] As warranted by widespread criticism of the SW-OLS procedure (Derksen and Keselman, 1992).

Technique	Dependence	Estimation	Hyperparameters	Criterion
RWD	temporal	MLE	–	–
ARIMA	temporal	MLE	p_i, q_i $(i = 1, 2, \ldots, N)$	BIC
DPLM*	temporal	GMM	p	significance
VAR	temporal and cross-sectional	LASSO	p	CV-MSE
STARIMA*	spatio-temporal	KF	p, λ_ℓ $(\ell = 1, 2, \ldots, p)$, q, μ_ℓ $(\ell = 1, 2, \ldots, q)$	BIC
SDPLM*	spatio-temporal	QML	weights $\in$ {contiguity, inverse-distance}, time lag $\in \{0, 1\}$, space-time lag $\in \{0, 1\}$, Lee–Yu transf. $\in \{0, 1\}$	quasi-BIC
ESTF*	spatio-temporal	SW-OLS or LASSO	weights $\in$ {contiguity, inverse-distance}, spec $\in$ {contemporaneous, lagged}, selection $\in$ {SW-OLS, LASSO}	BIC or DR

Table 4.1: Properties of the various techniques (*: not used previously in this context)

Given the time series k_{it} for each country from the LC model, univariate unit root and second-generation panel unit root tests (Hurlin and Mignon, 2007) may help determine their orders of integration. The latter account for cross-sectional dependence, which is central to our approach. Specifically, for ARIMA we use Augmented Dickey–Fuller (ADF) tests to determine country-specific orders d_i such that the differenced series have no unit roots. For all forecasting techniques other than RWD and ARIMA, we use the Covariate-Augmented Dickey–Fuller (CADF) test (Costantini and Lupi, 2013), a second-generation panel unit root test, to determine a uniform order of integration $d_i = d$ $(i = 1, 2, \ldots, N)$ across countries.

For the LL model, following Li and Lee (2005), we recommend forecasting the common trend K_t using an RWD process estimated by MLE. For the country-specific k_{it}, we set $d_i = 0$ and $d = 0$ across all countries and forecasting techniques to ensure coherence.

In the remainder of our paper, we will use the word "model" for the LC and LL models, and "technique" for the forecasting techniques presented in this subsection.[4]

4.4 Robust selection[5]

We propose dividing the dataset into a training set, consisting of the initial T_{train} years of available data, and a test set, comprising the remaining T_{test} years, with $T = T_{\text{train}} + T_{\text{test}}$, thereby spanning the entire observation period. Additionally, because model performance may be sensitive to the location of the train–test split, we propose a robust selection procedure involving repeated estimation across multiple training horizons $T_{\text{train}}^{(h)}$, with $T_{\text{test}}^{(h)} = T - T_{\text{train}}^{(h)}$ $(h = 1, 2, \ldots, H)$.

Forecast accuracy of forecasting technique j under model ℓ (LC or LL in our case) in country i given train–test split h is then evaluated using the mean squared error (MSE) of the forecasted logarithmic mortality rates:

$$\text{MSE}_{i,j,\ell}^{(h)} \triangleq \frac{1}{XT_{\text{test}}^{(h)}} \sum_{x=1}^{X} \sum_{t=T_{\text{train}}^{(h)}+1}^{T} \left(\ln m_{ixt} - \ln \hat{m}_{ixt}^{(j,\ell)} \right)^2,$$

where $\hat{m}_{(\cdot)}^{(j,\ell)}$ denotes mortality rates forecasted using technique j with its optimal hyperparameters under model ℓ.

As it is preferable to assess model performance relative to the best-performing procedure rather than against a perfect fit or another hypothetical benchmark, for each split, we define the underperformance scores

$$\text{U}_{i,j,\ell}^{(h)} \triangleq \text{MSE}_{i,j,\ell}^{(h)} - \min_{j,\ell} \text{MSE}_{i,j,\ell}^{(h)},$$

which measure how much worse forecasting technique j under model ℓ performs for country i and training horizon h, relative to the best-performing combination.

Model underperformance can then be assessed independently from the location

[4] Under this terminology, DPLM and SDPLM are techniques, despite what their names suggest.

[5] We avoid the term "model selection," as "selection of model–technique combinations" would be more accurate but unnecessarily cumbersome.

of the train–test split by measuring average or maximum underperformance (AU or MU) scores across the splits, with the maximum criterion being more conservative and reflecting the minimax principle from decision theory:

$$\mathrm{AU}_{i,j,\ell} \triangleq \frac{1}{H}\sum_{h=1}^{H} \mathrm{U}^{(h)}_{i,j,\ell},$$
$$\mathrm{MU}_{i,j,\ell} \triangleq \max_{h=1,2,\ldots,H} \mathrm{U}^{(h)}_{i,j,\ell}.$$

These measures can then be summarized by their medians across all countries, enabling the robust comparison of model–technique combinations:

$$\mathrm{MAU}_{j,\ell} \triangleq \operatorname*{median}_{i=1,2,\ldots,N} (\mathrm{AU}_{i,j,\ell}),$$
$$\mathrm{MMU}_{j,\ell} \triangleq \operatorname*{median}_{i=1,2,\ldots,N} (\mathrm{MU}_{i,j,\ell})$$

MAU captures median underperformance relative to the best-performing combination, while MMU reflects median worst-case underperformance across train–test splits.

4.5 Stochastic forecasts

Stochastic forecasts may be generated by incorporating some or all of the following four sources of uncertainty (Cairns, 2000):

1. *Measurement uncertainty* arises from the random error terms ε_{ixt} in the LC and LL log-mortality equations. It captures fluctuations in observed log-mortality rates that are not explained by the systematic components of the models.

2. *Process uncertainty* arises from the inherent randomness in the future evolution of the stochastic time series components k_{it} of the models.

3. *Parameter uncertainty* results from potential errors in the estimation of the model parameters a_{ix}, b_{ix}, k_{it}, B_x, and K_t based on finite historical data.

4. *Model uncertainty* reflects the possibility that the chosen model structure (e.g., LC–RWD) may be misspecified, or that alternative models could better represent the underlying mortality dynamics.

Following the approach recommended by Brouhns et al. (2005), we address parameter uncertainty by generating synthetic training datasets of bootstrapped death counts $\{D_{ixt}^{(r)}\}_{r=1}^{R}$ drawn independently from Poisson distributions with means $\lambda_{ixt} \triangleq D_{ixt}$, where R is the number of bootstrap samples. This yields bootstrapped estimates $a_{ix}^{(r)}$, $b_{ix}^{(r)}$, and $k_{it}^{(r)}$ separately for the LC and LL models, and $B_x^{(r)}$ and $K_t^{(r)}$ for the LL model.

We incorporate process uncertainty via resampling with replacement from the past residuals of the fitted time series model for the given country:[6]

$$\Delta^{(d_i)} k_{i,T_{\text{train}}+t}^{(r)} = \Psi_{it}^{(r)} + \xi_{it}^{(r)} \quad (t \geq 1,\ r = 1, 2, \ldots, R),$$

and additionally, for the LL model, we apply

$$\Delta K_{T_{\text{train}}+t}^{(r)} = \Theta^{(r)} + \eta_t^{(r)} \quad (t \geq 1,\ r = 1, 2, \ldots, R),$$

where $\Psi_{it}^{(r)} \triangleq \mathbb{E}_{T_{\text{train}}}\left(\Delta^{(d_i)} k_{i,T_{\text{train}}+t}^{(r)}\right)$ with $\mathbb{E}_{T_{\text{train}}}(\cdot)$ denoting a conditional expectation[7] based on bootstrapped samples up to T_{train}, $\Theta^{(r)}$ is the drift term of the RWD process of the common mortality index K_t, and $\xi_{it}^{(r)}$ and $\eta_t^{(r)}$ are randomly drawn with replacement from training-set residuals of $\Delta^{(d_i)} k_{it}$ and ΔK_t $(t = 1, 2, \ldots, T_{\text{train}})$. We draw from the residuals rather than from normal distributions, as the normality assumption may not hold and methods outside the ARIMA family do not rely on it.

We account for measurement uncertainty by adding bootstrapped residuals $\varepsilon_{i,x,T_{\text{train}}+t}^{(r)}$ drawn from past values $\{\hat{\varepsilon}_{ixt}\}_{t=1}^{T_{\text{train}}}$ to the systematic components of the log-mortality equations (with $B_x^{(r)} \triangleq K_t^{(r)} \triangleq 0$ in the case of the LC model):

$$\ln m_{i,x,T_{\text{train}}+t}^{(r)} \triangleq \ln m_{i,x,T_{\text{train}}}^{(r)} + b_{ix}^{(r)}\left(k_{i,T_{\text{train}}+t}^{(r)} - k_{i,T_{\text{train}}}^{(r)}\right) + B_x^{(r)}\left(K_{T_{\text{train}}+t}^{(r)} - K_{T_{\text{train}}}^{(r)}\right) + \varepsilon_{i,x,T_{\text{train}}+t}^{(r)}.$$

[6] We suppress the dependence on the model–technique combination to simplify the notation. Past residuals are computed from the original parameter estimates and not the bootstrapped ones.

[7] We are avoiding the sigma-algebra notation.

Model uncertainty could be accounted for within a fully Bayesian framework; however, we leave this aspect outside the scope of the present study, as subjective priors would introduce a new layer of uncertainty, and model averaging with BIC-derived weights is not feasible as some forecasting techniques we have proposed are not estimated by maximum likelihood. Instead, we apply post hoc adjustment and ensembling to mitigate this limitation.

To reduce sampling bias, we apply mean calibration, i.e., a pointwise shift to the bootstrapped trajectories so that their sample mean equals the expected trajectory $\hat{m}_{i,x,T_{\text{train}}+t}$ computed under the given model–technique combination at the end of the training period:

$$\ln \tilde{m}^{(r)}_{i,x,T_{\text{train}}+t} \triangleq \ln m^{(r)}_{i,x,T_{\text{train}}+t} + \ln \hat{m}_{i,x,T_{\text{train}}+t} - \frac{1}{R}\sum_{r=1}^{R} \ln m^{(r)}_{i,x,T_{\text{train}}+t}.$$

Then we construct bootstrapped prediction intervals with confidence level $1-\tau$ for the log-mortality rates $\ln m_{i,x,T_{\text{train}}+t}$ as

$$\left(\pi^{(L,\tau)}_{i,x,T_{\text{train}}+t}, \pi^{(U,\tau)}_{i,x,T_{\text{train}}+t}\right) \triangleq \left(Q_{\frac{\tau}{2}}\left(\{\ln \tilde{m}^{(r)}_{i,x,T_{\text{train}}+t}\}_{r=1}^{R}\right),\ Q_{1-\frac{\tau}{2}}\left(\{\ln \tilde{m}^{(r)}_{i,x,T_{\text{train}}+t}\}_{r=1}^{R}\right)\right),$$

where $Q_\tau(\cdot)$ denotes the empirical τ-quantile of the resampled estimates.

To correct potentially inaccurate interval coverage (e.g., due to model uncertainty or structural breaks), we apply a post hoc scaling adjustment by selecting the smallest scaling factor ν_t for which the empirical coverage probability on the validation set is at least the nominal level $1-\tau$, for each model–technique combination and forecast horizon $t = 1, 2, \ldots, T_{\text{val}}$:

$$\frac{1}{N\,|\mathcal{X}|}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}} 1_{\left(\pi^{*(L,\tau)}_{i,x,T_{\text{train}}+t} \le \ln m_{i,x,T_{\text{train}}+t} \le \pi^{*(U,\tau)}_{i,x,T_{\text{train}}+t}\right)} \ge 1-\tau,$$

where $1_{(\cdot)}$ are indicator functions, and

$$\left(\pi^{*(L,\tau)}_{i,x,T_{\text{train}}+t}, \pi^{*(U,\tau)}_{i,x,T_{\text{train}}+t}\right) \triangleq \left(\pi^{(L,\tau)}_{i,x,T_{\text{train}}+t} - \frac{\nu_t - 1}{2}\left(\pi^{(U,\tau)}_{i,x,T_{\text{train}}+t} - \pi^{(L,\tau)}_{i,x,T_{\text{train}}+t}\right), \pi^{(U,\tau)}_{i,x,T_{\text{train}}+t} + \frac{\nu_t - 1}{2}\left(\pi^{(U,\tau)}_{i,x,T_{\text{train}}+t} - \pi^{(L,\tau)}_{i,x,T_{\text{train}}+t}\right)\right).$$

We assess the quality of these prediction intervals on a separate evaluation period of length T_{eval}, with the original test period partitioned as $T_{\text{val}} + T_{\text{eval}} = T_{\text{test}}$, using average interval scores (Gneiting and Raftery, 2007) by country, computed as

$$\begin{aligned}\text{AIS}_{i,1-\tau} \triangleq \frac{1}{|\mathcal{X}|\, T_{\text{eval}}} \sum_{x\in\mathcal{X}} \sum_{t=1}^{T_{\text{eval}}} \Bigg[& \pi^{*(U,\tau)}_{i,x,T_{\text{train}}+T_{\text{val}}+t} - \pi^{*(L,\tau)}_{i,x,T_{\text{train}}+T_{\text{val}}+t} + \\ & + \frac{2}{\tau}\max\left(\pi^{*(L,\tau)}_{i,x,T_{\text{train}}+T_{\text{val}}+t} - \ln m_{i,x,T_{\text{train}}+T_{\text{val}}+t}, 0\right) + \\ & + \frac{2}{\tau}\max\left(\ln m_{i,x,T_{\text{train}}+T_{\text{val}}+t} - \pi^{*(U,\tau)}_{i,x,T_{\text{train}}+T_{\text{val}}+t}, 0\right)\Bigg],\end{aligned}$$

where we sample from the first $T_{\text{train}} + T_{\text{val}}$ years, simulate for the remaining T_{test} years, use mean calibration to expected trajectories estimated at the end of the validation period, and apply the scaling factors ν_t previously estimated on the validation period.

Finally, we compute median average interval scores across countries as a robust summary of prediction interval quality across model–technique combinations:

$$\text{MAIS}_{1-\tau} \triangleq \underset{i=1,2,\ldots,N}{\text{median}}\, \text{AIS}_{i,1-\tau},$$

and assess out-of-sample coverage using empirical coverage probabilities:

$$\text{ECP}_{1-\tau} \triangleq \frac{1}{N\,|\mathcal{X}|\,T_{\text{eval}}} \sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{t=1}^{T_{\text{eval}}} 1_{\left(\pi^{*(L,\tau)}_{i,x,T_{\text{train}}+T_{\text{val}}+t} \le \ln m_{i,x,T_{\text{train}}+T_{\text{val}}+t} \le \pi^{*(U,\tau)}_{i,x,T_{\text{train}}+T_{\text{val}}+t}\right)}.$$

This procedure could be made even more robust by using multiple train–test splits, following the analogy of the robust selection pipeline described in Subsection 4.4.

4.6 Empirical analysis

We used the R programming language (R Core Team, 2025) for our computations.[8] Figure 4.1 illustrates the steps of the analysis, from data collection to robust selection and the evaluation of the prediction intervals.

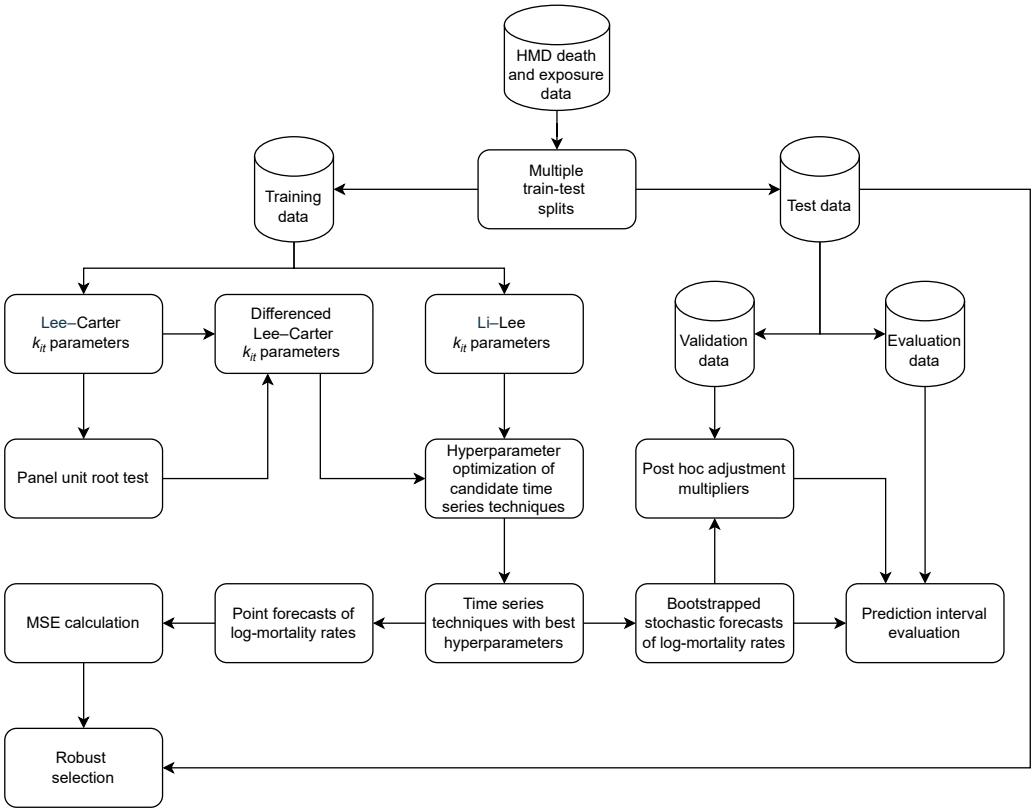

Figure 4.1: Analysis workflow

4.6.1 Data collection and preparation

We obtained our dataset from the Human Mortality Database (HMD, 2025). We collected unisex data on the number of deaths and exposures in all European countries for which data were available for the $T = 60$ years between 1960 and

[8] And in particular, the packages `StMoMo` (Villegas et al., 2015), `MultiMoMo` (Robben, 2021), `punitroots` (Kleiber and Lupi, 2011), `forecast` (Hyndman and Khandakar, 2008), `sparsevar` (Vazzoler, 2021), `plm` (Croissant and Millo, 2008), `starma` (Cheysson, 2016), and `SDPDmod` (Simonovska, 2025).

2019, categorized by country, age, and calendar year, and excluded the remaining countries. We did not use data for ages above 99 years due to missing data and other potential data quality issues. Consequently, our dataset comprised the $N = 22$ countries depicted in Figure 4.2.

Figure 4.2: Map of the countries included in the analysis with their ISO-2 codes

We used data only up to 2019 to avoid testing on the years of the COVID-19 pandemic, which would have led to a bias toward models that tend to overestimate mortality. Assuming that COVID-19 was a one-off event, the pandemic years could be included after subtracting COVID-19 deaths from total counts or by interpolating data for the years of the pandemic using rates from neighboring years. However, a significant problem with the subtraction approach is that criteria for recording the cause of death and the extent of testing varied widely by country, while the second approach is not yet viable as data for years following the end of the pandemic are not yet available for some countries in the Human Mortality

Database. Due to these constraints, we opted to perform our analysis on this truncated period.

By default, we used the first $T_{\text{train}} = 45$ years, up to 2004, as training data, the remaining $T_{\text{test}} = 15$ years as test data, and split the test period into a validation period of $T_{\text{val}} = 8$ years and an evaluation period of $T_{\text{eval}} = 7$ years.

4.6.2 Model estimation and interpretation

We estimated the parameters of the LC and LL models on data from all $N = 22$ countries. Table 4.2 displays the spatial and spatio-temporal Moran's I statistics of the differenced mortality indices Δk_{it} of the LC and LL models using contiguity and inverse-distance spatial weights.[9] These results indicate that there is significant spatial and spatio-temporal autocorrelation in mortality improvements in both the LC and LL models, which motivates the application of spatio-temporal forecasting techniques, and contiguity weighting yields stronger spatio-temporal autocorrelation than inverse-distance weighting.

	Contiguity weights			**Inverse-distance weights**		
Model	I	$I^{(C)}_{\text{space-time}}$	$I^{(L)}_{\text{space-time}}$	I	$I^{(C)}_{\text{space-time}}$	$I^{(L)}_{\text{space-time}}$
LC	0.709	0.097	-0.150	0.273	0.062	-0.147
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.000)
LL	0.774	0.039	-0.057	0.326	0.023	-0.090
	(0.000)	(0.093)	(0.005)	(0.000)	(0.273)	(0.000)

Table 4.2: Spatial and spatio-temporal Moran's I statistics of the differenced mortality indices Δk_{it} from the LC and LL models under alternative spatial weighting schemes (p-values in parentheses).

We performed LISA clustering at the 5% significance level on the means of Δk_{it} from the LC and LL models by country. We used contiguity weighting in this step, based on the findings reported in Table 4.2. LISA identified two spatial clusters for the LC model: France and Switzerland, and the Baltic states of Estonia,

[9] This descriptive step is not shown in Figure 4.1. Averages $\frac{1}{T-1}\sum_{t=2}^{T}\Delta k_{it}$ across time periods are used in the computation of spatial Moran's I. Contemporaneous spatio-temporal autocorrelation is not significant at the 5% level for the LL model, but it is significant at the 10% level using contiguity weighting. The inverse-distance weights are based on Haversine distances between country capitals.

Latvia, and Lithuania, both showing high mean differences in the mortality index. For the LL model, the first significant cluster consisted of France and Germany, and the second of the three Baltic states, again both with high mean differences.

The spatio-temporal eigenvectors with the largest absolute coefficients revealed clear spatial patterns. In the LC model, they separated the United Kingdom and Ireland, as well as the Scandinavian countries of Finland, Norway, and Sweden, from the rest of Europe under both Stepwise and LASSO selections. In the LL model, they again isolated these same clusters under LASSO, while under Stepwise the three Baltic states were separated from the rest of Europe. This has a natural geographic interpretation, as the separated groups of countries are divided by bodies of water from the rest of the countries in the analysis.

4.6.3 Forecasting techniques and hyperparameter optimization

We assessed the stationarity of the LC mortality indices using the CADF test, which yielded the p-values 0.847 on the original panel data and 0.000 after differencing, indicating the successful removal of unit roots, and $d = 1$ as the uniform order of integration.

For both the LC and LL models, we estimated the parameters and optimized the hyperparameters of the time series forecasting techniques as we outlined in Table 4.1. In all cases, the contiguity spatial weights matrix outperformed the version constructed from inverse distances between capital cities, in line with the findings from Table 4.2.

4.6.4 Point forecasts

We forecasted the country-specific mortality indices of the LC and LL models and the log-mortality rates using all forecasting techniques presented in Subsection 4.3, each with its optimal set of hyperparameters, using observed jump-off rates to enhance forecast accuracy, as recommended by Lee and Miller (2001), and applied an RWD for the common mortality index of the LL model.[10] The first row of

[10]Since the SW-OLS and LASSO methods rely on different selection criteria and are not directly comparable, we applied an unweighted average ensemble of the best SW-OLS and LASSO specifications for forecasting in both the LC and LL cases.

Figure 4.3 illustrates the LC model predictions for Norway across selected ages, and the second row shows the corresponding coherent LL forecasts.[11]

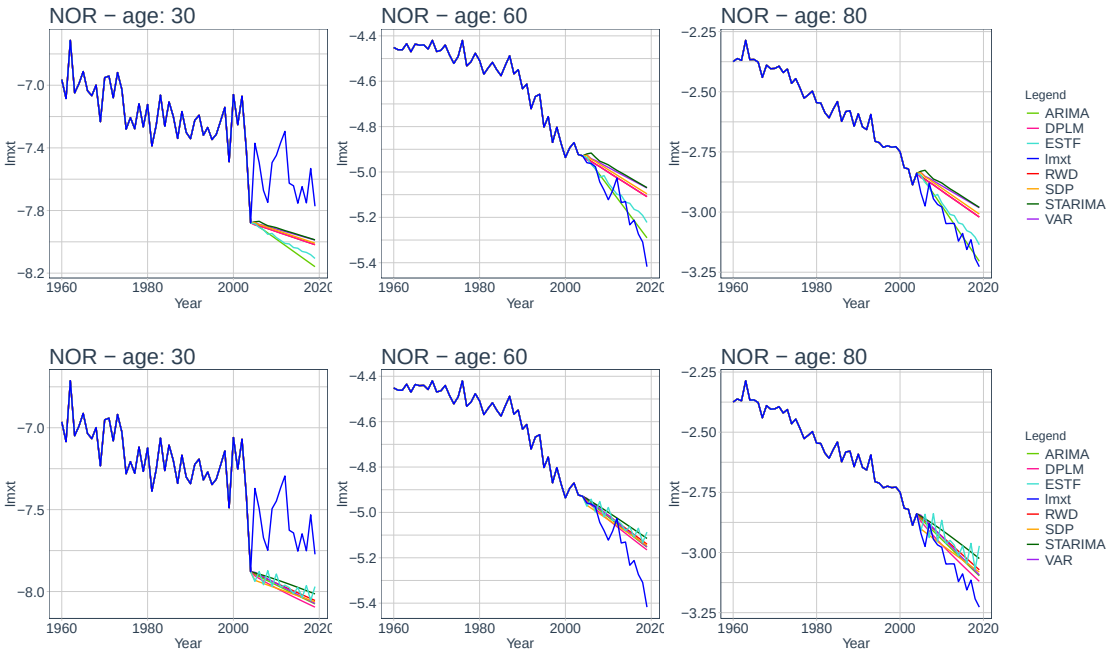

Figure 4.3: Point projections of log-mortality for selected ages in Norway from the LC (first row) and LL (second row) models.

4.6.5 Robust selection

Following forecasting, we applied the robust selection procedure described in Subsection 4.4, using the years 2002 through 2008 as the endpoints of the training period (i.e., $T_{\text{train}} \in \{42, 43, 44, 45, 46, 47, 48\}$) to evaluate the predictive performance of the combinations of the LC and LL models and the seven forecasting techniques and the two ensembles. The numbers and ISO-2 codes of the countries in which each technique performs best under each model, based on AU and MU, are summarized in Table 4.3.[12]

[11] Ensemble forecasts are not displayed to avoid visual clutter, but can be interpreted as (trimmed) average paths.

[12] In the case of the LL model, RWD, ARIMA, and STARIMA reduce to RW, ARMA, and STARMA, respectively, since no differencing is applied to ensure coherence.

Within the LC framework, LC–ESTF emerges as the strongest overall performer, with 6 and 7 wins under AU and MU, respectively. LC–SDPLM also performs well, achieving 4 wins under both criteria. Overall, the spatio-temporal techniques account for 17 and 14 wins under AU and MU, respectively, out of the 22 populations.

Within the LL model, LL–DPLM stands out with 6 and 5 wins under AU and MU, respectively, making it the best performer across criteria. LL–STARIMA follows with 3 and 5 wins. Spatio-temporal techniques are best in 9 and 11 countries under AU and MU, respectively—less frequently than in the LC framework. This is expected, as the common trend in the LL model already captures much of the cross-country dependence, and the spatio-temporal techniques can only leverage residual dependence beyond this.

Spatio-temporal techniques often outperform generic VAR as they rely on more parsimonious dependence structures.

	LC		**LL**	
Technique	**AU**	**MU**	**AU**	**MU**
RWD	1 (FR)	5 (AT,CH,FR,IE,PL)	3 (GB,IT,SE)	3 (CH,GB,IT)
ARIMA	1 (EE)	1 (EE)	1 (ES)	0
DPLM*	1 (HU)	1 (HU)	**6** (DK,EE,HU,IE,NO,PT)	**5** (DK,EE,IE,NO,PT)
VAR	2 (BE,IT)	1 (BE)	3 (CZ,LV,PL)	3 (CZ,HU,PL)
STARIMA*	1 (PL)	0	3 (DE,FR,NL)	5 (AT,BE,DE,FR,NL)
SDPLM*	4 (AT,DE,GB,PT)	4 (DE,GB,IT,PT)	0	0
ESTF*	**6** (CH,DK,ES,FI,NL,NO)	**7** (CZ,DK,ES,FI,LV,NL,NO)	2 (AT,FI)	2 (FI,SE)
ENS1*	4 (CZ,IE,LV,SK)	1 (SK)	1 (BE)	3 (ES,LT,LV)
ENS2*	2 (LT,SE)	2 (LT,SE)	3 (CH,LT,SK)	1 (SK)

Table 4.3: Number of wins by technique within each model, and ISO-2 codes of countries where they won, based on AU and MU. Asterisks denote newly proposed techniques in this context, and highest win counts by column are typeset in bold underlined font.

We provide another comprehensive summary of our results in Figure 4.4, where the red bars display MAU, the red and blue bars combined correspond to values of the more cautious MMU, and the left and right panels refer to the LC and LL models, respectively. This comparison is chiefly relevant if the same model–

technique combination is applied in all countries. Otherwise, country-specific combinations are clearly superior.

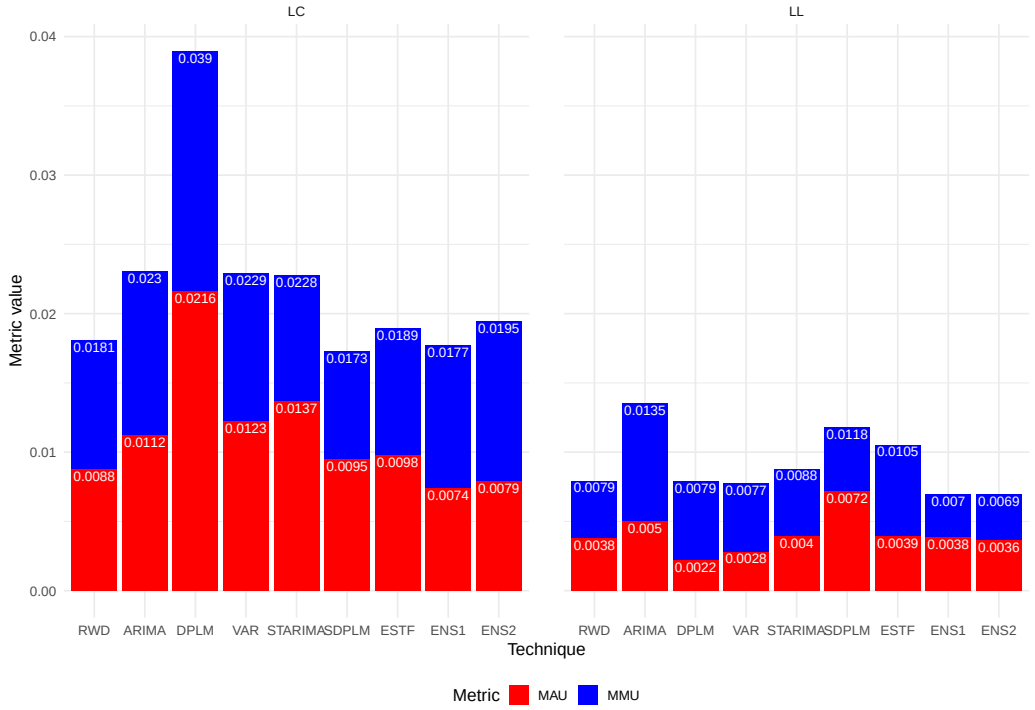

Figure 4.4: Comparison of MAU and MMU values across models and techniques (MMU is the sum of the heights of the red and blue columns)

As shown in Figure 4.4, the LL–DPLM model achieves the lowest overall MAU, but is slightly outperformed by the ensembles LL–ENS1 and LL–ENS2 in terms of the more conservative MMU. Within the LC model, the ensembles LC–ENS1 and LC–ENS2 achieve the lowest MAU, while LC–SDPLM is the best in terms of MMU.

Notably, ensembles achieve excellent overall MAU and MMU, and LL-based forecasts outperform LC-based ones, as they incorporate information on the common trend distilled from a larger set of data in addition to country-specific patterns, while remaining coherent.

Overall, on the set of countries, ages and calendar years under consideration, LL–DPLM, LC–ESTF, LL–VAR and LL–STARIMA emerge as strong performers, with notable variation by country. The widely used RWD and ARIMA

specifications are clearly outperformed by other procedures in both the LC and LL frameworks.

4.6.6 Stochastic forecasts

We performed the procedure presented in Subsection 4.5 using $T_{\text{train}} = 45$, $T_{\text{val}} = 8$, and $T_{\text{eval}} = 7$ years, $R = 100$ bootstrap samples, and a confidence level of $1 - \tau = 0.9$. For the sake of illustration, the first and second rows of Figure 4.5 display mean-calibrated, post hoc adjusted, bootstrapped 90% prediction intervals of log-mortality rates for selected ages in Norway, based on the LC–ESTF and LL–ESTF frameworks.[13]

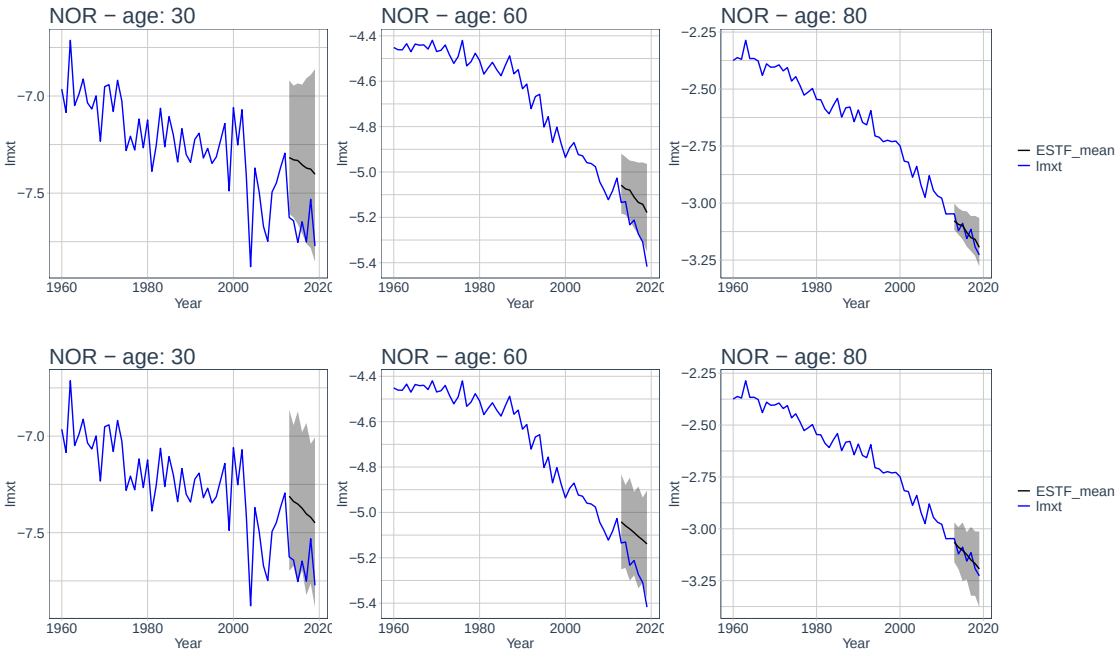

Figure 4.5: Bootstrapped 90% prediction intervals for selected ages in Norway from the LC–ESTF (first row) and LL–ESTF (second row) frameworks.

As ENS1 and ENS2 lack their own mortality index series, we computed their prediction intervals as quantiles of the means and trimmed means, respectively, of the bootstrapped samples of the log-mortality rates from the seven individual

[13]Prediction uncertainty at age 30 is dominated by measurement uncertainty, which does not accumulate over time.

techniques, and applied all subsequent modeling steps in the same way as to the original techniques.

Table 4.4 reports the number and list of countries in which each technique achieves the lowest average 90% interval score under the LC and LL models. Under LC, SDPLM performs best, winning in six countries, followed by STARIMA, DPLM and ENS1, while the classical benchmarks RWD and ARIMA win in one country each. Under LL, DPLM attains the highest number of wins (seven), closely followed by ESTF (six), whereas ARIMA wins in only two countries and RWD in none.

Technique	**LC**	**LL**
RWD	1 (IE)	0
ARIMA	1 (LT)	2 (ES, IT)
DPLM*	3 (DK, PL, SK)	**7** (BE, CH, CZ, HU, NO, PL, PT)
VAR	1 (DE)	2 (IE, NL)
STARIMA*	4 (EE, FI, NO, PT)	1 (GB)
SDPLM*	**6** (AT, FR, GB, HU, IT, LV)	1 (EE)
ESTF*	2 (BE, CZ)	6 (AT, DK, FI, LT, LV, SK)
ENS1*	3 (CH, ES, SE)	1 (SE)
ENS2*	1 (NL)	2 (DE, FR)

Table 4.4: Number of countries in which each forecasting technique yields the lowest average 90% interval score, and ISO-2 codes of countries where they won. Asterisks denote newly proposed techniques, and highest win counts by model are typeset in bold underlined font.

Table 4.5 reports median average 90% interval scores and empirical coverage probabilities across forecasting techniques. Under the LC model, ENS1 achieves the lowest median interval score, closely followed by ENS2 and SDPLM, while classical benchmarks exhibit higher scores. Under the LL model, ESTF attains the lowest MAIS, with DPLM performing similarly well. Empirical coverage probabilities are generally close to the nominal level across methods, while LL-based frameworks tend to slightly overcover.

	LC		LL	
Technique	$\text{MAIS}_{0.90}$	$\text{ECP}_{0.90}$	$\text{MAIS}_{0.90}$	$\text{ECP}_{0.90}$
RWD	0.997	0.864	1.029	0.929
ARIMA	0.946	0.845	0.898	0.888
DPLM*	0.878	0.894	0.827	0.913
VAR	0.865	0.900	0.896	0.909
STARIMA*	0.892	0.913	0.883	0.908
SDPLM*	0.865	0.896	0.931	0.922
ESTF*	0.897	0.878	**0.827**	0.934
ENS1*	**0.845**	0.893	0.885	0.916
ENS2*	0.848	0.897	0.873	0.899

Table 4.5: Median average 90% interval scores and empirical coverage probabilities by forecasting technique under the LC and LL models. Asterisks denote newly proposed techniques, and lowest MAIS by model are typeset in bold underlined font.

4.6.7 Robustness check

To assess the robustness of our forecasts, we compared period life expectancy at birth at the end of the test period in 2019, computed using the various forecasting techniques estimated on the training period 1960–2005, for all $N = 22$ countries and both the LC and LL models. This choice was motivated by four considerations: (i) life expectancy at birth is an easily interpretable measure with an intuitive scale; (ii) it summarizes mortality across all ages into a single indicator, allowing for a parsimonious comparison; (iii) by the end of the test period, forecasts from different techniques are expected to have diverged sufficiently to allow for an informative comparison; and (iv) unlike cohort life expectancies, it does not require re-estimating the models and techniques on the full dataset and forecasting an additional 100 years, which would be more speculative as well as computationally demanding.

By careful individual examination of the projected mortality trajectories by country, model, and technique, we identified a few anomalies. In the Baltic states of Estonia, Latvia, and Lithuania, the LC mortality index series k_{it} increases over time, accompanied by negative values of b_{ix} at childhood and some other ages and positive values at middle ages. This pattern captures the idiosyncratic mortality experience of these countries, where middle-aged mortality stagnated or

worsened during the late Soviet period (Krúmiņš and Zvidriņš, 1992), while other age groups experienced improving mortality. STARIMA introduces decreases into k_{it} over several periods, which leads to projected increases in child mortality and decreases in middle-aged mortality in the forecasts for all three countries. A similar inversion is observed for ESTF in Estonia and Latvia. As this behavior is clearly a modeling artifact, we exclude LC–STARIMA for all three countries and LC–ESTF for Estonia and Latvia. This inversion does not affect the robust selection or the evaluation of prediction intervals presented in Subsections 4.6.5 and 4.6.6, as these techniques were not selected for the affected countries and the reported medians are robust to outliers.

The results are displayed in Figure 4.6, which shows little variation across techniques. The standard deviations across techniques, computed separately for each country and model, range from 0.01 to 1.06 years, with a median of 0.44 years, corroborating the robustness of our forecasts.

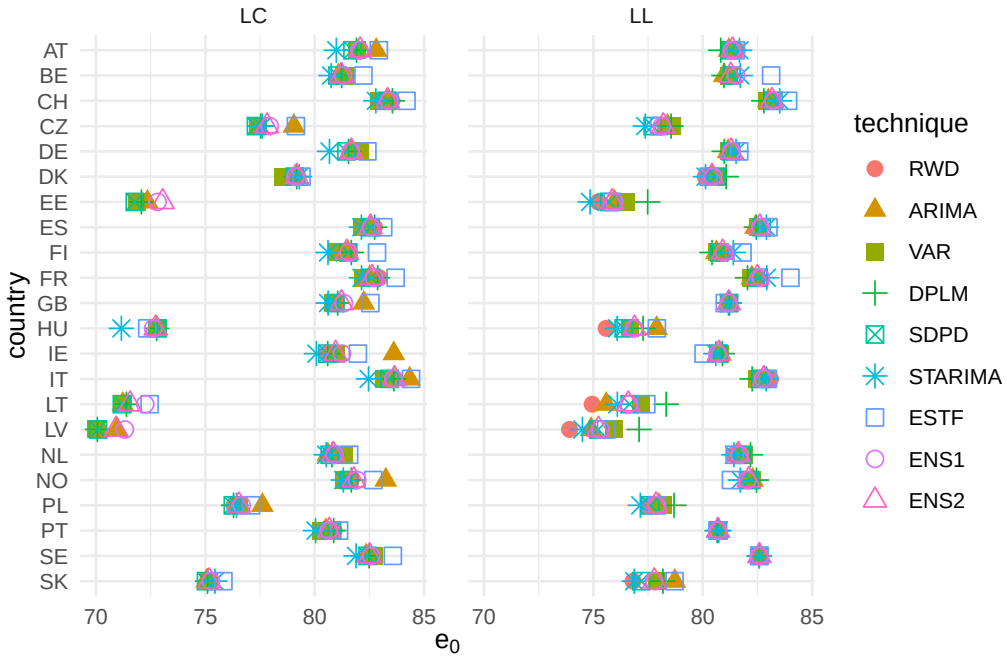

Figure 4.6: Comparison of projected period life expectancy at birth in 2019 across countries, models, and techniques

4.7 Conclusion

Our paper introduces a suite of multipopulation forecasting techniques for the Lee–Carter and Li–Lee country-specific mortality indices. Traditionally, most authors apply RWD or ARIMA models. By contrast, we have also explored various spatial and panel econometric techniques. We applied these to mortality data from 22 countries, spanning the years 1960 to 2019 and covering ages 0 to 99. We estimated both the Lee–Carter and Li–Lee models on training periods ending in years ranging from 2001 through 2007, and forecasted their mortality indices using various techniques, with coherent forecasts in the Li–Lee frameworks. We then analyzed the forecast accuracy based on our proposed robust selection procedure using test sets spanning the remaining years through 2019.

Based on multiple robust metrics (number of wins by average and maximum underperformance relative to the best model–technique combination, as well as median average underperformance and maximum average underperformance), the combinations of the Li–Lee and the Dynamic Panel Linear Models, the Lee–Carter model and the Eigenvector Spatio-Temporal Filter, the Li–Lee model and Vector Autoregression, and the Li–Lee model and Spatio-Temporal ARIMA emerge as particularly strong performers in point forecasts, with notable variation by country. Out of these frameworks, we are the first to propose all but the third one in the literature to the best of our knowledge.

We propose a mean-calibrated, post hoc adjusted bootstrapping procedure to generate prediction intervals with correct nominal coverage. In interval forecasting, the newly introduced techniques clearly outperform classical benchmarks, with the Li–Lee model combined with the Eigenvector Spatio-Temporal Filter or the Dynamic Panel Linear Model achieving the lowest median average interval scores.

We examined period life expectancies at birth in 2019 across the various model–technique combinations, and found that the forecasts were sufficiently robust.

In summary, our results provide evidence that spatial and panel econometric techniques for forecasting the Lee–Carter and Li–Lee mortality indices can be superior to more established approaches, and applying them may yield measurable improvements in accuracy in several populations.

Our study is limited in that it only considers the Lee–Carter and Li–Lee models as the two most prominent general and coherent multipopulation stochastic mortality models, and the Random Walk with Drift for the common mortality index in the Li–Lee model, and does not delve into Bayesian techniques or more sophisticated ensembles for forecasting the mortality indices.

Despite the built-in robustness, the results may change if the set of countries, ages, and calendar years considered in the analysis is modified.

A further limitation is the use of geographical distance, as countries' mortality experiences may be more closely aligned through cultural proximity. This could be accounted for using spatial weights matrices based on Hofstede's cultural dimensions theory (Hofstede, 1984), among other possible spatial weighting schemes.

Additionally, we have omitted the years of the COVID-19 pandemic from the testing period out of necessity, and our analysis is currently limited by the absence of data for the post-pandemic years.

Finally, it would be worthwhile to apply the outlined methodology to more granular data—e.g., at the levels of counties, states, or other subnational units—where even more meaningful spatial relationships may emerge.

The proposed techniques lend themselves to practical applications by actuaries and demographers, facilitating more accurate forecasting of mortality patterns and enhancing the design of demographic projections, pension systems, and life insurance products.

Chapter 5

Case study: Mortality forecasting in Visegrád Group countries using recurrent neural networks

This chapter is based on the author's previously published work Szentkereszti and Vékás (2024). A similar case study can also be found in Szentkereszti and Vékás (2022).

5.1 Introduction

In this analysis, we apply recurrent neural networks, which have been used with great success in many fields in recent years, to predict age-specific mortality rates for ages 18 to 99 years for the Czech Republic, Poland, Hungary and Slovakia (Visegrád Group countries) between 1970 and 2019, separately for women and men. We investigate the popular Recurrent Neural Network, long-short term memory and Gated Recurrent Unit architectures, and place special emphasis on the optimization of the hyperparameters of the networks by applying cross-validation and splitting the base period into learning and testing subsets. We compare our predictions with those of both the classical Lee-Carter and the coherent Li-Lee multipopulation models in terms of accuracy, and determine which procedures are able to generate the most reliable projections for each country, gender and age

group. Our models can be applied in practice by life, pension and health actuaries as well as demographers.

5.2 Modeling process

We downloaded mortality data for the Visegrád Group countries from the HMD (2025), where annual numbers of deaths and central exposures are available by age and sex for the four countries. We used data for calendar years from 1970 to 2019, focusing on the adult population aged 18–99. We chose this age group because life insurance products are rarely offered for children, and the improvement in child mortality differs significantly from that of other age groups. For those over 99, there are very few data available, which would make forecasts uncertain.

Data preparation and modeling were carried out using the R programming language (R Core Team, 2025). The Lee–Carter (LC) models were implemented using the StMoMo package (Villegas et al., 2015), the Li–Lee (LL) models with the MultiMoMo package (Robben, 2021), and the neural networks using the keras package (Allaire and Chollet, 2022).

The selected period was split into training and testing sets. The training set covered the period from 1970 to 2009, while the testing set included the period after 2009, as we considered the latter most relevant for forecasting purposes. To optimize the hyperparameters of the machine learning methods, we applied fivefold simple cross-validation. After identifying the optimal hyperparameters, the models with the best-performing hyperparameter configurations were trained on the entire training set.

For the LC model and the machine learning methods, the data were treated separately by country and sex, so we built models for each country–sex pair. In the case of the LL model, due to its multipopulation nature, we did not split the four populations by sex, resulting in two multipopulation models.

For the neural network models, the logarithmic mortality rates were normalized, and a grid search was performed to identify the best model. The sets of hyperparameters were narrowed down based on the problem's scale as well as professional and literature-based considerations, as follows:

- k: lags of the target variable, $k \in \{3, 5, 8\}$.
- b: batch size, $b \in \{64, 128, 256\}$.
- u: number of neurons, $u \in \{5, 6, 7\}$, and $u \in \{10, 12, 14\}$ for the RNN.

For all networks, the target variable was the logarithmic mortality rate corresponding to a given age and calendar year, while the input values consisted solely of the lagged values of the target variable.

The number of epochs was set to $e = 2000$, and early stopping was applied with a patience hyperparameter of $p = 100$ to reduce computation time. Since the problems were relatively small in scale, we used only single-layer networks. The recurrent layer employed the widely used hyperbolic tangent activation function, while the output layer of the RNN used the standard identity activation function. Training was performed using the popular Adaptive Moment Estimation (Kingma and Ba, 2014) stochastic optimization with a learning rate of $\lambda = 0.001$ and the mean squared error (MSE) loss function.

Due to the stochastic nature of training, we built 50 models for each final parameter set: from the 50 training runs, we selected the best-performing model and also computed the average of all model forecasts, which we refer to as the ensemble model. The methodology for neural network forecasts followed the approach presented by Richman and Wüthrich (2019) in the international literature.

5.3 Results

The performance of the models on the female data is summarized in Table 5.1. The table contains aggregated data for the testing set, taking all ages into account.

Figure 5.1 summarizes the errors of the 50–50 trained models on the testing period, by model and country. The dashed red line indicates the performance of the LC model, while the dotted purple line represents that of the LL model.

Figure 5.1 shows that, in the case of Slovakia, the classical methods performed well compared to the neural networks, and there is little difference between the LC and LL models. For Hungary, the LL model was generally much better than the LC model. In the Czech Republic, the opposite is observed: the LC model performed

Country	LC	LL	LSTM B	LSTM E	GRU B	GRU E	RNN B	RNN E
CZE	2,21	3,24	1,94	2,07	1,96	2,02	1,98	2,20
HUN	7,40	4,70	3,31	3,91	3,08	3,78	3,46	4,08
POL	1,58	1,91	1,16	1,29	1,06	1,25	1,22	1,46
SVK	4,91	4,63	4,27	4,46	4,34	4,82	4,22	4,57

Table 5.1: Forecasting errors by model for different countries.
B = best, E = ensemble

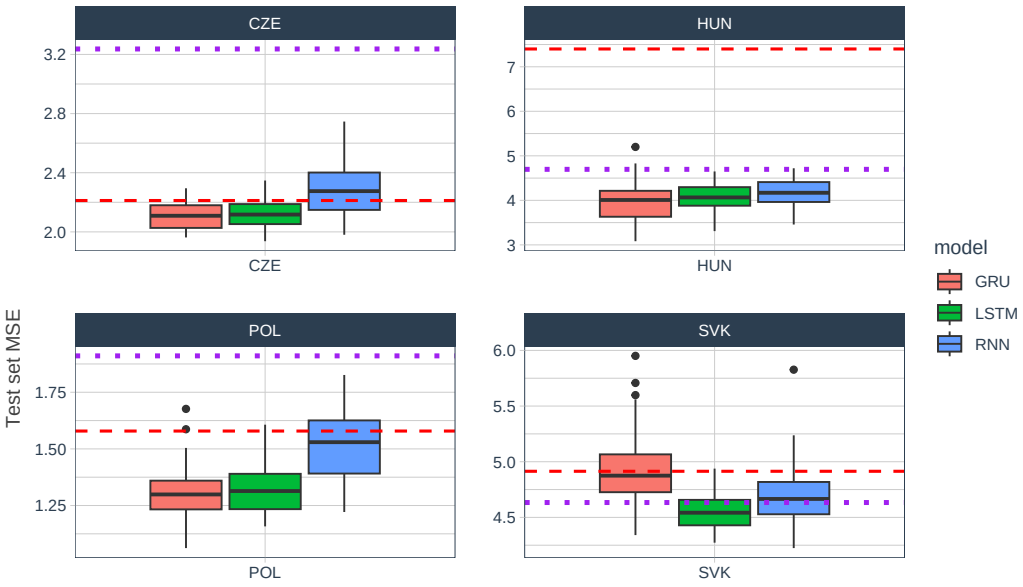

Figure 5.1: Test set errors for the female data by country and model

much better than the LL model. In Poland, there is again little difference between the two classical models, but the machine learning methods proved superior. Thus, on the female data, we see that the performance of the machine learning models surpasses that of the LC and LL models, and this holds true even when considering not only the best training run but also the more robust ensemble forecasts.

When analyzing the results, it is important to examine the strengths and weaknesses of the models in detail, in addition to the aggregate error metrics. Figure 5.2

shows, by age and country, which model performed best for the female data, i.e., which model had the lowest age-specific MSE. For robustness, the results of the ensemble models are used here. The results indicate that the LL model performed very well in the older age groups for the female data, except in Hungary, whereas the machine learning models achieved the best performance in the middle-aged and younger populations. The LC model was the best-performing model for the older Hungarian population.

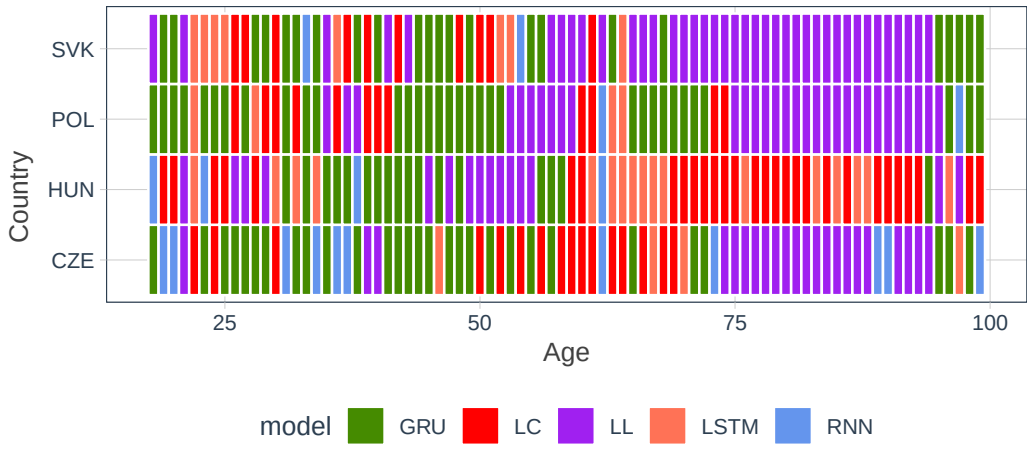

Figure 5.2: Best-performing model by age and country on the female test data

Table 5.2 summarizes the results shown in Figure 5.2. Overall, for the female data, the GRU ensemble performed best in most age groups in the Czech Republic and Poland, the LC model performed best in Hungary, and the LL model performed best in Slovakia. However, the ensemble models also include the results of poorly performing training runs, and as seen in Table 5.1, using the best training run yields considerably better performance for the machine learning models.

In addition to the counts, it is important to show how the magnitude of the errors compares across models. Therefore, we plotted the mean squared errors by age on the testing set for each model and country.

For the male data, the modeling process was identical to that applied to the female data. The corresponding results are presented in Table 5.3 and Figure 5.3. In the Czech Republic and Poland, the GRU performed best, while in Hungary the

Model	CZE	HUN	POL	SVK
LC	16	29	12	10
LL	22	15	31	39
RNN ensemble	11	4	2	2
LSTM ensemble	4	17	4	8
GRU ensemble	29	17	33	23

Table 5.2: Number of age groups in which each model performed best on the female data, by country

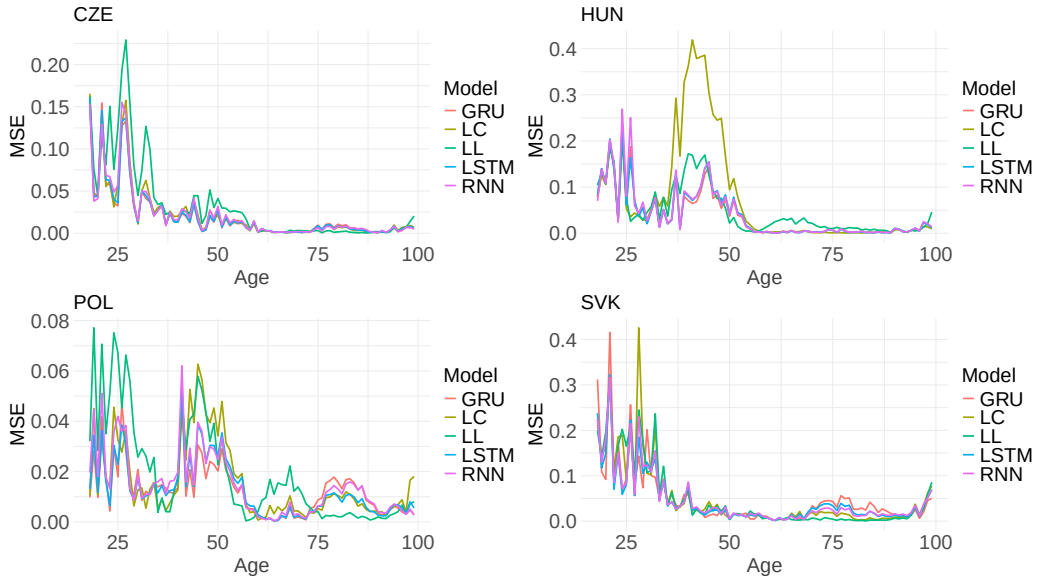

Figure 5.3: Age-specific MSE values on the test set for the female data

LSTM achieved the best results, and in Slovakia, the Li–Lee model most accurately predicted male mortality rates.

In many cases, the results are similar to those observed for the female data. In the Czech Republic, the LC model was more accurate than the LL model, whereas for Hungary, the LL model significantly outperformed the LC. Overall, in both countries, the neural network methods provide the best forecasts. In Poland, the competition was closer: the LL model approached the performance of the best neural network forecasts and even outperformed the ensemble models. In Slovakia, the classical models performed better than this type of neural network approach. From a performance and robustness perspective, a practical compromise could be

to average the results of the ten best training runs.

Country	LC	LL	LSTM B	LSTM E	GRU B	GRU E	RNN B	RNN E
CZE	1,83	2,62	1,31	1,49	1,28	1,52	1,35	1,57
HUN	17,75	6,11	1,77	2,99	2,11	3,91	2,07	3,44
POL	1,39	1,17	1,10	1,41	0,97	1,39	1,05	1,38
SVK	3,07	2,40	3,20	4,02	3,30	4,08	3,16	4,15

Table 5.3: Forecasting errors by model for different countries.
B = best, E = ensemble

Figure 5.4: Test set errors for the male data by country and model

Figure 5.5 shows, by age and country, which model performed best for the male data (for robustness, as in Figure 5.4, the results of the ensemble models are displayed). For Slovakia and Poland, the LL model performs very well above ages 50 and 40, respectively, while in Hungary, the LSTM ensemble performs best between ages 35 and 50, and the GRU ensemble performs best between ages 70

and 80. In the Czech Republic, the GRU ensemble shows outstanding performance between ages 40 and 65.

Figure 5.5: Best-performing model by age and country on the male test data

Table 5.4 quantifies the results shown in Figure 5.5. It can be seen that for the male data, the LL model is frequently the best-performing model in Slovakia and Poland. In Hungary, the LSTM ensemble models proved to be the best. For the Czech Republic, the GRU ensemble model performs best for most ages, as it works very well for the middle-aged population, while for the older age groups, the performance is more mixed. It is important to note that these are the results of the ensemble models, and as seen in Table 5.3, when considering the best training runs, the performance of the neural networks is considerably more favorable.

Table 5.4: Number of age groups in which each model performed best on the male data, by country

Model	CZE	HUN	POL	SVK
LC	17	4	12	15
LL	17	21	57	56
RNN ensemble	17	14	5	3
LSTM ensemble	4	26	5	2
GRU ensemble	27	17	3	6

Similar to the female data, here we also show how the models performed

relative to each other by age, based on the mean squared error.

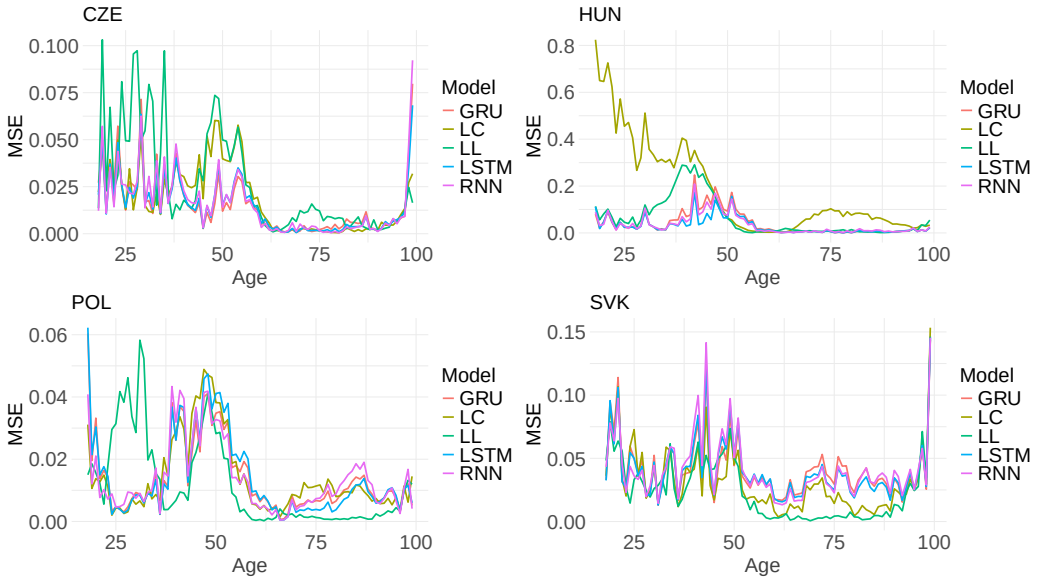

Figure 5.6: Age-specific MSE values on the test set for the male data

It is important to note that, since neural networks are not stochastic models, forecasts produced using them do not include confidence intervals, although the mean squared errors provide an approximate measure of the accuracy of the point estimates.

The results do not imply that neural networks should generally be preferred to classical mortality models. Their relative advantage varies considerably across countries, sexes, and age groups. The LC and LL models remain competitive, and in several cases outperform the neural networks, particularly at older ages and for some Slovak and Polish populations. Moreover, classical stochastic mortality models are more parsimonious and interpretable and provide a natural framework for quantifying forecast uncertainty. Neural networks may therefore be preferable when they deliver a sufficiently large and robust improvement in out-of-sample point forecast accuracy, whereas classical models remain attractive when interpretability, parameter uncertainty, and prediction intervals are important.

Beyond the testing set, we also produced longer-term forecasts to examine the stability of the neural network methods. Additionally, we calculated the single net premiums of life annuities paying 1 unit per year at age 65 using different methods,

providing a practically relevant example. For the neural network models, we selected the best model for each country–sex combination from those presented above. For the life annuities, we set the maximum age at 99 and the technical interest rate at 3%. The prediction was also started from 2010, corresponding to the boundary between the training and testing sets.

Figure 5.7 shows the mortality rates from age 65 to 99 obtained using different approaches for each sex and country. The reference point is the mortality rates from the 2009 life table at the boundary between the training and testing sets.

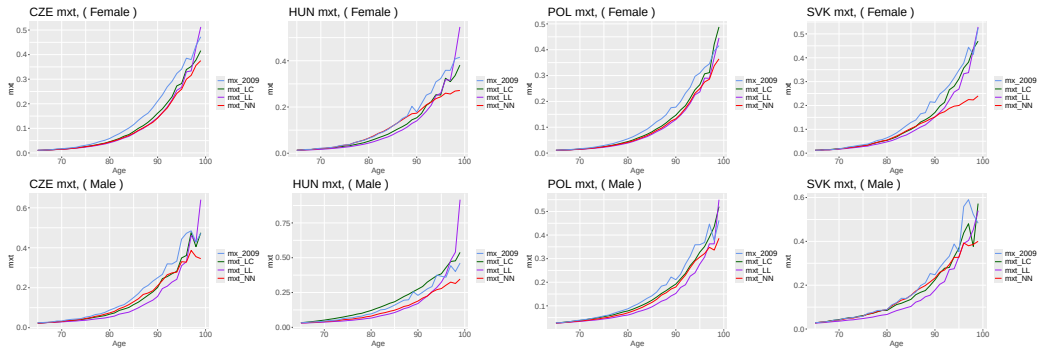

Figure 5.7: Mortality rate forecasts by different methods, by country and gender

Table 5.5 shows the single net premiums for life annuities paying 1 unit per year starting at age 65, based on female data.

Table 5.5: Single net premium factors of life annuities based on female mortality data

Forecasting method	CZE	HUN	POL	SVK
Static (2009)	14.23	13.78	14.43	13.82
LC	14.97	14.40	15.07	14.29
LL	15.23	14.92	15.43	14.78
NN	15.14	13.87	15.23	14.40

In general, it can be stated that the LC model produces the lowest premium after the static approach, the neural network models produce intermediate premiums, and the LL model produces the highest. Hungary is the only exception, due to the fact (as seen in the figure above) that the neural network estimates higher mortality rates around ages 85–90 than the classical methods.

We performed the same calculations for the male data and the results can be seen in Table 5.6.

Table 5.6: Single net premium factors of life annuities based on male mortality data

Forecasting method	CZE	HUN	POL	SVK
Static (2009)	12.10	11.17	11.77	11.40
LC	12.82	10.48	12.04	11.52
LL	13.43	12.31	12.80	12.46
NN	12.61	11.71	12.39	11.52

5.4 Conclusion

In this study, we analyzed and forecasted mortality data for the Visegrád countries using the traditionally widely applied Lee–Carter (LC) model, the multipopulation Li–Lee (LL) model, and recurrent neural networks (RNNs). The mortality patterns of the countries studied differ from those of Western European countries due to characteristics observed during the socialist era, making it important to analyze them separately. In our analysis, we developed direct RNN, LSTM, and GRU models for the Visegrád countries, optimized their hyperparameters, selected the best models, and then forecasted mortality rates on a predefined testing set. Finally, we compared the results of the models both overall and by age group for all four countries and for both sexes.

Based on the aggregated results, the best-trained neural network models consistently achieved the highest performance on the testing set. To ensure the robustness of the forecasts, we also constructed ensemble models. Across the eight populations studied, with the exception of the Slovak and Polish male data, the ensemble neural networks achieved the best overall results. For the two exceptions mentioned, the LL model had the lowest error. When examining the models by age, it was observed that in older age groups, the LL model was most often the best, whereas the neural networks achieved the best results mainly in middle-aged and younger populations. However, our results cannot be directly compared to those published in the international literature. First, other authors often used different

models and approaches. Second, the Visegrád countries or specifically Hungary have not previously been studied in this way.

Compared to the study by Szentkereszti and Vékás (2022), our investigation extends beyond Hungary to the group of Visegrád countries, which show similar historical mortality trends. We tested several recurrent neural network architectures: in addition to the previously used LSTM networks, we also present results from RNN and GRU models. Moreover, to obtain more robust results, in addition to the best-performing network on the testing set, we present the average forecasts of models constructed with the same hyperparameters. We also produced longer-term forecasts using the best-trained neural networks and examined the net premiums of life annuities calculated using static and various dynamic forecasting methods.

Future research could investigate the mortality data of Central and Visegrád Group countries using other machine learning methods reported in the literature and extend our analysis to additional countries. Instead of separate models, a unified model encompassing data from all studied countries could be developed. It may also be important to examine the medium- and long-term forecasts produced by neural networks. Using our results, some assumptions underlying demographic forecasts and pension calculations could be reconsidered, and life insurance premium calculations could be made more accurate compared to using static (period) or LC-based cohort life tables.

Chapter 6

Coherent mortality forecasting by feedforward neural networks

This chapter is based on ongoing joint work by the author, Vékás, and Wang.

6.1 Introduction

Mortality forecasting plays a vital role in demography, actuarial science, and public policy. Reliable projections underpin pension sustainability analyses, life insurance pricing, and the planning of long-term healthcare and social security systems. While the seminal Lee–Carter model (Lee and Carter, 1992) and its numerous extensions and alternatives have substantially advanced the field, these traditional frameworks rely on predefined structures that may limit their ability to capture complex age–period–cohort interactions and nonlinear relationships inherent in real-world mortality dynamics. A further challenge arises when forecasting mortality for multiple related populations, such as males and females or different regions of the same country. In such cases, coherence—the requirement that long-term mortality ratios between populations converge to finite constants over time—is crucial for ensuring plausible projections. The Li–Lee model (Li and Lee, 2005) provides a coherent framework by introducing common and population-specific factors, yet its functional form restricts flexibility.

Recent advances in machine learning, and in particular deep neural networks,

offer a compelling alternative. Feedforward neural networks (FNNs) can approximate highly nonlinear mappings and discover latent structures directly from data without imposing strong parametric assumptions. However, despite their flexibility, little attention has been paid to the theoretical question of whether neural-network-based mortality forecasts can satisfy the coherence property that underpins multipopulation models.

This paper develops a theoretical framework for coherent multipopulation mortality forecasting using deep FNNs. We formally prove that such forecasts preserve coherence when activation functions satisfy certain conditions. We further classify activation functions by their asymptotic coherence behavior into hypercoherent, convergent, weakly coherent, and incoherent types, providing a unified taxonomy. We present an empirical analysis on data from England & Wales using a train–validation–test split to demonstrate that weakly coherent activations such as the Swish function achieve a favorable balance of coherence, flexibility, and out-of-sample forecasting performance, and can significantly outperform the Li–Lee benchmark model across multiple age groups relevant in actuarial practice.

The remainder of this chapter is structured as follows. Section 6.2 discusses the relevant gap in the literature. Section 6.3 presents the theoretical framework and sufficient conditions for coherence in FNN-based multipopulation mortality forecasting. Section 6.4 presents the empirical results and a comparison with the Li–Lee benchmark model.

6.2 Gap in the literature

The review by Zheng et al. (2025) identifies coherent multipopulation forecasting, quantifying prediction uncertainty, and interpretability as key challenges in contemporary mortality forecasting. They mention two studies (Bravo, 2021 and Perla and Scognamiglio, 2023) that explicitly address coherence. While they describe the model of Bravo (2021) as producing coherent forecasts, the latter in fact generates smooth and consistent projections without explicitly investigating or ensuring the formal criterion of coherence, which will be discussed in Section 6.3.

The model of Perla and Scognamiglio (2023) does yield locally coherent forecasts while retaining the Lee–Carter functional form, and these forecasts are

also globally coherent if only one single cluster is postulated.

Additionally, De Mori et al. (2025) employ multi-task FNNs for multipopulation mortality forecasting without enforcing coherence. They note among future research directions that penalization could be introduced into their estimation procedure to encourage a degree of coherence.

In summary, there appears to be no deep FNN architecture in the literature that directly models mortality rates without imposing a Lee–Carter structure, as in Richman and Wüthrich (2021) and De Mori et al. (2025), and at the same time achieves coherent forecasts. Our work aims to fill this gap.

6.3 Methodology

In Subsection 6.3.1, we introduce a taxonomy of coherence in multipopulation mortality forecasts (MMFs) to facilitate the discussion in subsequent parts of the paper, followed in Subsection 6.3.2 by the presentation of our benchmark, the augmented common factor model of Li and Lee (2005), and an outline of the simple deep FNN architecture that directly outputs log-mortality rates in Subsection 6.3.3.

6.3.1 A taxonomy of coherence

Let $\mathcal{P} = \{1, 2, \ldots, N\}$ be a set of populations, $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$ a set of ages, $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ a set of calendar periods, $m_{x,t}^{(p)}$ the central mortality rate of population $p \in \mathcal{P}$ for age $x \in \mathcal{X}$ in period $t \in \mathcal{T}$, and let $\hat{m}_{x,t_n+h}^{(p)}$ $(h = 1, 2, \ldots)$ denote its model-implied point forecast based on mortality observations up to time t_n.

Definition 1 (Coherent MMF)**.** Following Li and Lee (2005), an MMF is *coherent* if the ratio of the long-term forecasts for a given age and any two populations converges to a positive constant over time, i.e.,

$$\lim_{h \to +\infty} \frac{\hat{m}_{x,t_n+h}^{(i)}}{\hat{m}_{x,t_n+h}^{(j)}} = c_x^{(i,j)} \in \mathbb{R}_{++} \tag{6.1}$$

for each $x \in \mathcal{X}$ and $i, j \in \mathcal{P}$.

Building on this basic definition, we introduce three additional coherence concepts to facilitate our remaining discussion.

Definition 2 (Convergent MMF)**.** An MMF is *convergent* if (6.1) holds with

$$c_x^{(i,j)} = 1 \tag{6.2}$$

for each $x \in \mathcal{X}$ and $i, j \in \mathcal{P}$, implying that age-specific rates are asymptotically identical in different populations.

While assuming coherence is plausible, postulating convergence is generally too restrictive, since, for example, women's mortality is consistently lower than men's in practice.

Definition 3 (Hypercoherent MMF)**.** An MMF is *hypercoherent* if $\exists \tau \geq 0$ such that

$$\frac{\hat{m}_{x,t_n+h}^{(i)}}{\hat{m}_{x,t_n+h}^{(j)}} = c_x^{(i,j)} \in \mathbb{R}_{++} \tag{6.3}$$

for any $h \geq \tau$, each $x \in \mathcal{X}$ and $i, j \in \mathcal{P}$, i.e., mortality rates for the same age and different populations are proportional (and log-mortality curves are parallel) beyond a time threshold.

Assuming hypercoherence is likewise too restrictive, as it precludes changes in ratios of mortality rates beyond the threshold, whereas in practice such ratios tend to constantly evolve over time (e.g., the life expectancy gap between China and Western countries has narrowed in practice).

Definition 4 (Weakly coherent MMF)**.** An MMF is *weakly coherent* if it is coherent, but neither convergent, nor hypercoherent.

For the reasons discussed above, we argue that postulating weak coherence is the most defensible assumption in practice.

Definition 5 (Incoherent MMF)**.** An MMF is *incoherent* if (6.1) does not hold.

Figure 6.1 visualizes the hierarchy of these concepts in a tree diagram. Convergent, hypercoherent, and weakly coherent are special cases of coherence, while the

first two overlap in the special case where forecasts become identical across populations beyond a time threshold. Additionally, Figure 6.2 displays one example of each of the four types.

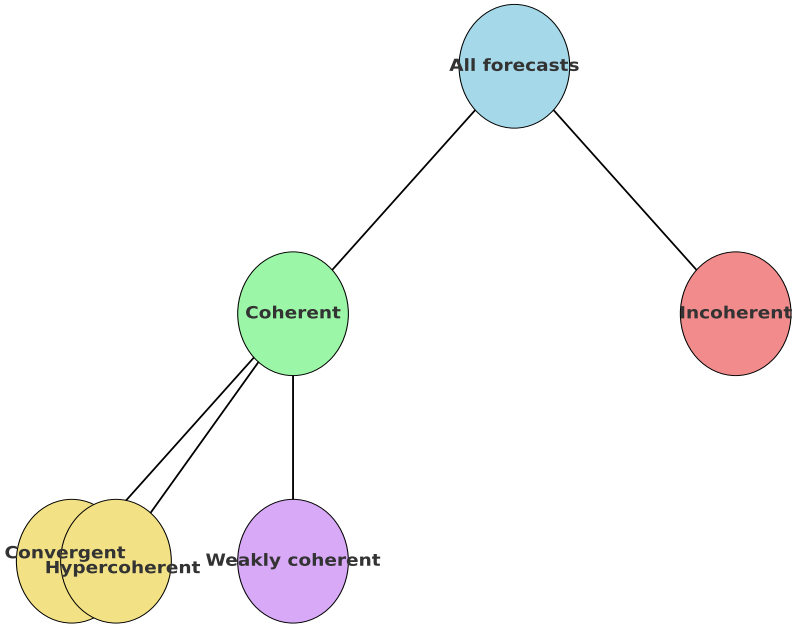

Figure 6.1: Hierarchy of MMFs by coherence type

Each of the listed coherence properties can apply globally, as in the previous definitions, or locally (Perla and Scognamiglio, 2023), i.e., within subgroups of a partition of populations:

Definition 6 (Local coherence properties)**.** Let $\{\mathcal{P}_j\}_{j=1}^{q}$ be a partition of $\mathcal{P}$ such that

$$\bigcup_{j=1}^{q} \mathcal{P}_j = \mathcal{P}, \qquad \mathcal{P}_{j_1} \cap \mathcal{P}_{j_2} = \emptyset \quad \text{for } j_1 \neq j_2, \qquad \text{and} \quad |\mathcal{P}_j| \geq 2 \quad \text{for all } j = 1, 2, \ldots, q.$$

For a property $C \in \{\text{coherent}, \text{convergent}, \text{hypercoherent}, \text{weakly coherent}\}$, an MMF is *locally* C with respect to the partition if its restriction to each subgroup $\mathcal{P}_j$ has property C.

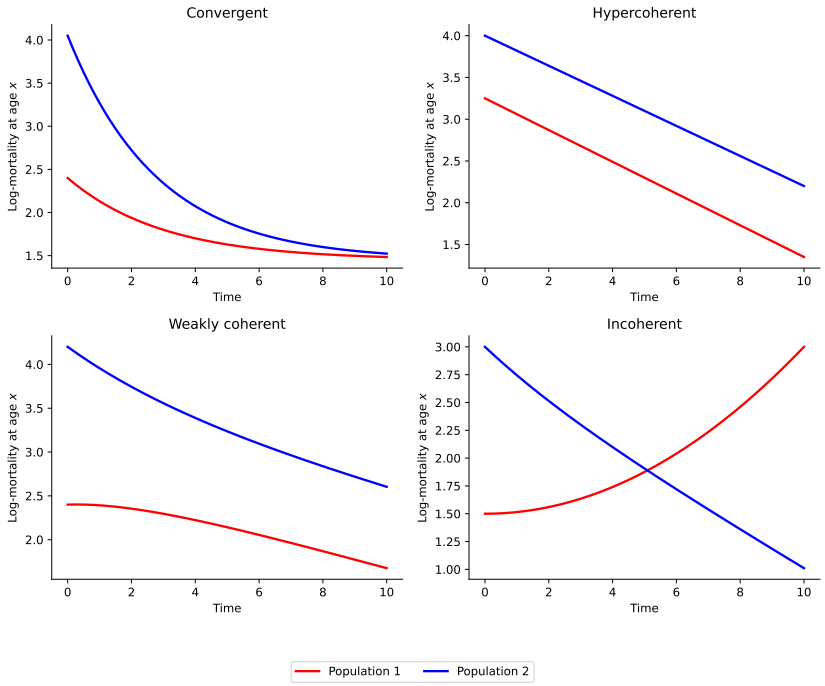

Figure 6.2: Examples of coherence types

6.3.2 The Li–Lee model

To remain consistent with the notation of the underlying joint paper, we use $\alpha_x^{(p)}$, $\beta_x^{(p)}$, and $\kappa_t^{(p)}$ in this chapter; these correspond to $a_{p,x}$, $b_{p,x}$, and $k_{p,t}$, respectively, in the notation used in Chapter 3.

We will use the model of Li and Lee (2005) (LL), specified as

$$\log(m_{x,t}^{(p)}) = \alpha_x^{(p)} + B_x K_t + \beta_x^{(p)} \kappa_t^{(p)} + \varepsilon_{x,t}^{(p)}$$

for $x \in \mathcal{X}$, $t \in \mathcal{T}$, and $p \in \mathcal{P}$, as our benchmark.

While K_t is typically projected using a random walk with drift, there are multiple possible specifications for the population-specific mortality index series $\kappa_t^{(p)}$,

each corresponding to different coherence properties presented in Subsection 6.3.1:

$$\begin{aligned}
\text{Random walk (RW):} \quad & \kappa_t^{(p)} = \kappa_{t-1}^{(p)} + \xi_t^{(p)}, \\
\text{First-order autoregressive process (AR(1)):} \quad & \kappa_t^{(p)} = \gamma^{(p)} + \rho^{(p)} \kappa_{t-1}^{(p)} + \xi_t^{(p)}, \\
\text{Random walk with drift (RWD):} \quad & \kappa_t^{(p)} = \gamma^{(p)} + \kappa_{t-1}^{(p)} + \xi_t^{(p)},
\end{aligned}$$

with $\xi_t^{(p)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_p^2)$.

Let

$$\hat{K}_{t_n+h} \triangleq \mathbb{E}_{t_n}(K_{t_n+h}), \qquad \hat{\kappa}_{t_n+h}^{(p)} \triangleq \mathbb{E}_{t_n}(\kappa_{t_n+h}^{(p)}),$$

where $\mathbb{E}_{t_n}$ denotes conditional expectation given mortality observations up to time t_n.[1] The model-implied point forecast is defined by[2]

$$\log\left(\hat{m}_{x,t_n+h}^{(p)}\right) = \alpha_x^{(p)} + B_x \hat{K}_{t_n+h} + \beta_x^{(p)} \hat{\kappa}_{t_n+h}^{(p)}.$$

Proposition 1 (Coherence properties of the Li–Lee model)**.** Consider the Li–Lee model with the population mortality indices $\kappa_t^{(p)}$ projected using the specifications above. Then:

(i) under the RW specification, the resulting MMF is hypercoherent;

(ii) under the AR(1) specification with $0 < |\rho^{(p)}| < 1$ for all $p \in \mathcal{P}$, let

$$A_x^{(p)} \triangleq \beta_x^{(p)} \left(\kappa_{t_n}^{(p)} - \frac{\gamma^{(p)}}{1 - \rho^{(p)}} \right).$$

Then the resulting MMF is weakly coherent, except if

$$\alpha_x^{(i)} - \alpha_x^{(j)} + \beta_x^{(i)} \frac{\gamma^{(i)}}{1 - \rho^{(i)}} - \beta_x^{(j)} \frac{\gamma^{(j)}}{1 - \rho^{(j)}} = 0$$

for all $i, j \in \mathcal{P}$ and $x \in \mathcal{X}$, in which case it is convergent, or if

$$A_x^{(i)} = A_x^{(j)}, \qquad A_x^{(i)} \left(\rho^{(i)} - \rho^{(j)} \right) = 0$$

[1] We are avoiding the sigma-algebra notation.

[2] If the conditional distribution of log mortality is symmetric, including under Gaussian innovations, this point forecast coincides with the conditional median of the mortality rate.

for all $i, j \in \mathcal{P}$ and $x \in \mathcal{X}$, in which case it is hypercoherent.

(iii) under the RWD specification, the resulting MMF is incoherent unless

$$\beta_x^{(i)}\gamma^{(i)} = \beta_x^{(j)}\gamma^{(j)}$$

for all $i, j \in \mathcal{P}$ and $x \in \mathcal{X}$, in which case it is hypercoherent.

Proof of Proposition 1.

(i) Since $\hat{\kappa}_{t_n+h}^{(p)} = \kappa_{t_n}^{(p)}$ for every $h \geq 1$,

$$\begin{aligned}\log(\hat{m}_{x,t_n+h}^{(i)}) - \log(\hat{m}_{x,t_n+h}^{(j)}) &= \alpha_x^{(i)} + B_x\hat{K}_{t_n+h} + \beta_x^{(i)}\hat{\kappa}_{t_n+h}^{(i)} - \left(\alpha_x^{(j)} + B_x\hat{K}_{t_n+h} + \beta_x^{(j)}\hat{\kappa}_{t_n+h}^{(j)}\right) \\ &= \alpha_x^{(i)} - \alpha_x^{(j)} + \beta_x^{(i)}\kappa_{t_n}^{(i)} - \beta_x^{(j)}\kappa_{t_n}^{(j)} = C_x^{(i,j)} \in \mathbb{R},\end{aligned}$$

thus the RW case results in a hypercoherent MMF.

(ii) In the AR(1) case with $0 < |\rho^{(p)}| < 1$,

$$\hat{\kappa}_{t_n+h}^{(p)} = \frac{\gamma^{(p)}}{1-\rho^{(p)}} + (\rho^{(p)})^h\left(\kappa_{t_n}^{(p)} - \frac{\gamma^{(p)}}{1-\rho^{(p)}}\right).$$

Hence

$$\begin{aligned}\log(\hat{m}_{x,t_n+h}^{(i)}) - \log(\hat{m}_{x,t_n+h}^{(j)}) = \alpha_x^{(i)} - \alpha_x^{(j)} + \beta_x^{(i)}\frac{\gamma^{(i)}}{1-\rho^{(i)}} - \beta_x^{(j)}\frac{\gamma^{(j)}}{1-\rho^{(j)}} \\ + A_x^{(i)}(\rho^{(i)})^h - A_x^{(j)}(\rho^{(j)})^h.\end{aligned}$$

Since $|\rho^{(p)}| < 1$, the last two terms converge to zero as $h \to +\infty$, and therefore

$$\lim_{h\to+\infty}\left[\log(\hat{m}_{x,t_n+h}^{(i)}) - \log(\hat{m}_{x,t_n+h}^{(j)})\right] = \alpha_x^{(i)} - \alpha_x^{(j)} + \beta_x^{(i)}\frac{\gamma^{(i)}}{1-\rho^{(i)}} - \beta_x^{(j)}\frac{\gamma^{(j)}}{1-\rho^{(j)}}.$$

Thus, the resulting MMF is coherent. It is convergent if this limit is zero for all $i, j \in \mathcal{P}$ and $x \in \mathcal{X}$.

It remains to characterize the hypercoherent case, which implies by definition that

$$A_x^{(i)}(\rho^{(i)})^h = A_x^{(j)}(\rho^{(j)})^h$$

for all sufficiently large h, all $i, j \in \mathcal{P}$, and all $x \in \mathcal{X}$. Because $0 < |\rho^{(p)}| < 1$, equality at two consecutive sufficiently large horizons implies

$$A_x^{(i)} = A_x^{(j)} \quad \text{and} \quad A_x^{(i)}\left(\rho^{(i)} - \rho^{(j)}\right) = 0.$$

Conversely, if these two conditions hold for all $i, j \in \mathcal{P}$ and $x \in \mathcal{X}$, then

$$A_x^{(i)}(\rho^{(i)})^h = A_x^{(j)}(\rho^{(j)})^h$$

for every h, so the MMF is hypercoherent.

Therefore, except for the convergent and hypercoherent cases identified above, the AR(1) specification yields a weakly coherent MMF.

(iii) Finally, in the RWD case,

$$\hat{\kappa}_{t_n+h}^{(p)} = \kappa_{t_n}^{(p)} + h\gamma^{(p)}.$$

Therefore,

$$\begin{aligned}\log(\hat{m}_{x,t_n+h}^{(i)}) - \log(\hat{m}_{x,t_n+h}^{(j)}) &= \alpha_x^{(i)} - \alpha_x^{(j)} + \beta_x^{(i)}\hat{\kappa}_{t_n+h}^{(i)} - \beta_x^{(j)}\hat{\kappa}_{t_n+h}^{(j)} \\ &= \alpha_x^{(i)} - \alpha_x^{(j)} + \beta_x^{(i)}(\kappa_{t_n}^{(i)} + h\gamma^{(i)}) - \beta_x^{(j)}(\kappa_{t_n}^{(j)} + h\gamma^{(j)}).\end{aligned}$$

The above expression fails to converge to a finite limit as $h \to \infty$ whenever

$$\beta_x^{(i)}\gamma^{(i)} \neq \beta_x^{(j)}\gamma^{(j)}$$

for some $x \in \mathcal{X}$ and $i, j \in \mathcal{P}$, resulting in an incoherent MMF, and is otherwise constant in h, resulting in a hypercoherent MMF.

□

6.3.3 A deep FNN for multipopulation mortality forecasting

In the remainder of our paper, we adopt the fully connected deep FNN architecture of Richman and Wüthrich (2021).[3] We model the negative log-mortality rates

$$y_{x,t}^{(p)} \triangleq -\log(m_{x,t}^{(p)})$$

as a function of age x, period t, and population p, without imposing a Lee–Carter functional form. The architecture is illustrated in Figure 6.3.

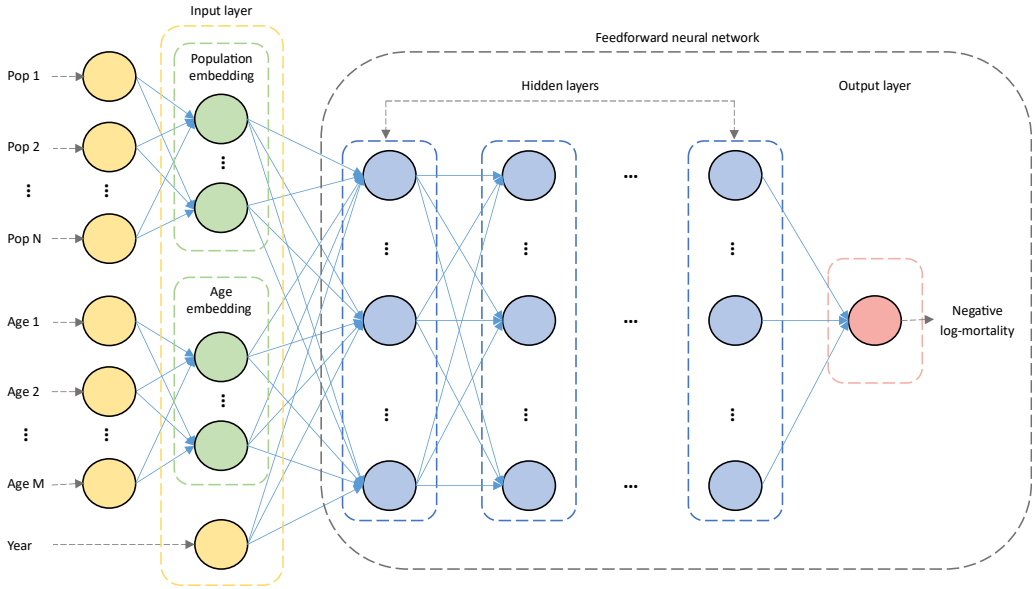

Figure 6.3: Architecture of the neural network (biases are not shown explicitly)

We treat calendar time t as a numerical covariate. Age and population are treated as categorical variables and represented through embedding layers. Specifically, for age we define

$$\mathbf{x} \triangleq g(x),$$

where $g : \mathcal{X} \to \mathbb{R}^{n_x}$ with $\mathcal{X} \triangleq \{x_1, \ldots, x_M\}$ maps each age category x_i to a

[3] As opposed to the authors, who model log-mortality, we use negative log-mortality rates to obtain an increasing function of calendar time t under mortality improvement. Additionally, the authors examine the addition of skip connections from the input layer directly into the output, which we omit for simplicity.

trainable embedding vector

$$g(x_i) = \mathbf{x}_i \in \mathbb{R}^{n_x}.$$

The vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ are learned parameters of the model, and n_x denotes the embedding dimension. The embedding for population can be dealt with similarly, and we use the notations $\mathbf{p}_j$ $(j = 1, \ldots, N)$ and n_p to represent the population embedding and its dimension.

We define

L: the depth of the neural network (including the output layer),

n_l: the number of nodes in the l-th layer with $n_L \triangleq 1$ $(l = 1, \ldots, L)$,

$\mathbf{W}^{(x)}$: $n_1 \times n_x$, the weight matrix for the age embedding in the input layer,

$\mathbf{w}_s^{(x)}$: $1 \times n_x$, the s-th row of $\mathbf{W}^{(x)}$ $(s = 1, \ldots, n_1)$,

$\mathbf{W}^{(p)}$: $n_1 \times n_p$, the weight matrix for the population embedding in the input layer,

$\mathbf{w}_s^{(p)}$: $1 \times n_p$, the s-th row of $\mathbf{W}^{(p)}$ $(s = 1, \ldots, n_1)$,

$\mathbf{w}^{(t)}$: $n_1 \times 1$, the weight vector for the calendar time t in the input layer,

$w_s^{(t)}$: the s-th element of $\mathbf{w}^{(t)}$ $(s = 1, \ldots, n_1)$,

$\mathbf{W}^{(l)}$: $n_l \times n_{l-1}$, the weight matrix of the l-th layer $(l = 2, \ldots, L)$,

$W_{rs}^{(l)}$: the (r, s)-th element of $\mathbf{W}^{(l)}$ $(r = 1, \ldots, n_l,\ s = 1, \ldots, n_{l-1})$,

$W_s^{(L)}$: the s-th element of the $1 \times n_{L-1}$ weight vector $\mathbf{W}^{(L)}$ $(s = 1, \ldots, n_{L-1})$,

$\mathbf{b}^{(l)}$: $n_l \times 1$, the bias vector of the l-th layer $(l = 1, \ldots, L)$,

$b_s^{(l)}$: the s-th element of $\mathbf{b}^{(l)}$ $(s = 1, \ldots, n_l)$, with $b^{(L)} \triangleq b_1^{(L)}$,

$u_{sj}^{(l)}(t)$ the pre-activation of the s-th neuron in layer l for population j and period t $(s = 1, \ldots, n_l,\ l = 1, \ldots, L,\ j \in \mathcal{P},\ t > 0)$,[4]

[4] As we will assume a fixed age, we suppress the dependence on age for simplicity.

$\boldsymbol{\phi}^{(l)} : \mathbb{R}^{n_l} \to \mathbb{R}^{n_l}$, the activation mapping of the l-th layer ($l = 1, \ldots, L-1$), obtained by applying the scalar activation function $\phi^{(l)} : \mathbb{R} \to \mathbb{R}$ componentwise,[5]

$\mathbf{h}^{(l)}(j,t)$: $n_l \times 1$, the output of the l-th layer for population j and period t ($l = 1, \ldots, L$, $j \in \mathcal{P}$, $t > 0$),

$h_s^{(l)}(j,t)$: the s-th element of $\mathbf{h}^{(l)}(j,t)$ ($s = 1, \ldots, n_l$), with $\hat{y}_{x_i,t}^{(j)} \triangleq h_1^{(L)}(j,t)$.

Definition 7 (Feedforward Mortality Network)**.** With the above notation, we define the Feedforward Mortality Network (FeMoNet) as follows:

$$\begin{aligned}
\mathbf{h}^{(1)}(j,t) &= \boldsymbol{\phi}^{(1)}(\mathbf{W}^{(x)}\mathbf{x}_i + \mathbf{W}^{(p)}\mathbf{p}_j + \mathbf{w}^{(t)}t + \mathbf{b}^{(1)}),\\
\mathbf{h}^{(l)}(j,t) &= \boldsymbol{\phi}^{(l)}(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)}(j,t) + \mathbf{b}^{(l)}) \quad (2 \le l \le L-1),\\
\hat{y}_{x_i,t}^{(j)} &= \sum_{s=1}^{n_{L-1}} W_s^{(L)} h_s^{(L-1)}(j,t) + b^{(L)}.
\end{aligned}$$

The network parameters are estimated by solving

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}} \sum_{p \in \mathcal{P}} F\left(y_{x,t}^{(p)}, \hat{y}_{x,t}^{(p)}(\boldsymbol{\theta})\right),$$

where $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ denotes a loss function—typically the squared error—and $\boldsymbol{\theta}$ collects all trainable parameters (embeddings, weights, and biases) in vectorized, concatenated form, while the notation $\hat{y}_{x,t}^{(p)}(\boldsymbol{\theta})$ expresses the dependence of the predictions on the parameters. The optimization is carried out using stochastic gradient methods. In the remainder of this section, the results hold regardless of the specific parameter values, and we therefore suppress the explicit dependence on $\boldsymbol{\theta}$ in the notation.

Theorem 1 (Sufficient condition of coherence)**.** The mortality forecast generated by FeMoNet has the coherence property if the activation functions $\phi^{(l)}$ ($l = 1, \ldots, L-1$) satisfy the following conditions:[6]

[5] We assume that the output layer is linear. In line with standard practice in applied deep learning, we assume a common activation function within each layer. The results extend to neuron-specific activation functions; the proof is analogous but technically more cumbersome and is omitted.

[6] We do not prescribe differentiability everywhere in Condition (i) because several popular

(i) $\phi^{(l)}(v)$ is continuous for all $v \in \mathbb{R}$ and differentiable for all $v \neq 0$;
(ii) there exist constants $d_+^{(l)} \geq 0$, $d_-^{(l)} \geq 0$, $e_+^{(l)}, e_-^{(l)} \in \mathbb{R}$ such that

$$\lim_{v \to +\infty} \left[\phi^{(l)}(v) - (d_+^{(l)} v + e_+^{(l)}) \right] = \lim_{v \to -\infty} \left[\phi^{(l)}(v) - (d_-^{(l)} v + e_-^{(l)}) \right] = 0$$

with $\lim_{v \to +\infty} \phi^{(l)\prime}(v) = d_+^{(l)}$ and $\lim_{v \to -\infty} \phi^{(l)\prime}(v) = d_-^{(l)}$ (it has nondecreasing linear asymptotes at both extremes).

Lemma 1 (Difference of activations)**.** Let $\phi : \mathbb{R} \to \mathbb{R}$ satisfy Conditions (i)–(ii) of Theorem 1. Let $f(t)$ and $g(t)$ be real-valued functions such that $\lim_{t \to +\infty} [f(t) - g(t)]$ exists and is finite. Assume moreover that either $f(t), g(t) \to +\infty$, or $f(t), g(t) \to -\infty$, or both converge in $\mathbb{R}$. Then the limit $\lim_{t \to +\infty} [\phi(f(t)) - \phi(g(t))]$ exists and is finite.

Proof of Lemma 1. If $f(t)$ and $g(t)$ converge in $\mathbb{R}$, the claim follows from continuity of ϕ. Otherwise, suppose $f(t), g(t) \to +\infty$ (the case $f(t), g(t) \to -\infty$ is analogous). For large enough t, the interval between $f(t)$ and $g(t)$ excludes 0. Hence by Condition (i) the Mean Value Theorem applies and yields a point $\xi(t)$ between $f(t)$ and $g(t)$ such that

$$\phi(f(t)) - \phi(g(t)) = \phi'(\xi(t)) \, [f(t) - g(t)].$$

Since $\xi(t) \to +\infty$, Condition (ii) implies $\phi'(\xi(t)) \to d_+$. As $f(t) - g(t)$ has a finite limit, the product converges, proving the claim. □

Proof of Theorem 1. Fix $x_i \in \mathcal{X}$ and $j, k \in \mathcal{P}$. We show that $\lim_{t \to +\infty} \left[\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)} \right]$ exists and is finite.

For each layer $l = 1, \ldots, L-1$ and neuron $s = 1, \ldots, n_l$, define the differences

$$\Delta h_s^{(l)}(t) \triangleq h_s^{(l)}(j,t) - h_s^{(l)}(k,t), \qquad \Delta u_s^{(l)}(t) \triangleq u_{sj}^{(l)}(t) - u_{sk}^{(l)}(t).$$

Step 1 (base layer: $l = 1$). From the input layer definition,

$$\Delta u_s^{(1)}(t) = \mathbf{w}_s^{(p)} (\mathbf{p}_j - \mathbf{p}_k), \tag{6.4}$$

activation functions (e.g., ReLU and ELU) are not differentiable at zero. The first part of Condition (ii) in itself does not imply the convergence of the derivatives in some oscillatory cases.

which is constant in t. Moreover $u_{sj}^{(1)}(t)$ is affine in t with slope $w_s^{(t)}$, hence $u_{sj}^{(1)}(t)$ and $u_{sk}^{(1)}(t)$ either both diverge to $+\infty$, both diverge to $-\infty$, or both remain constant. Applying Lemma 1 with $\phi = \phi^{(1)}$, $f(t) = u_{sj}^{(1)}(t)$, and $g(t) = u_{sk}^{(1)}(t)$, we conclude that $\lim_{t\to+\infty} \Delta h_s^{(1)}(t)$ exists and is finite for every $s = 1, \ldots, n_1$.

Step 2 (induction). Assume that for some $l \in \{1, \ldots, L-2\}$ the limits $\lim_{t\to+\infty} \Delta h_r^{(l)}(t)$ exist and are finite for all $r = 1, \ldots, n_l$. Fix any $s = 1, \ldots, n_{l+1}$. Using the layer definition and cancellation of the bias term,

$$\Delta u_s^{(l+1)}(t) = \sum_{r=1}^{n_l} W_{sr}^{(l+1)} \, \Delta h_r^{(l)}(t). \tag{6.5}$$

Taking $t \to +\infty$ yields that $\lim_{t\to+\infty} \Delta u_s^{(l+1)}(t)$ exists and is finite. Consequently, $u_{sj}^{(l+1)}(t)$ and $u_{sk}^{(l+1)}(t)$ share the same asymptotic behavior: they either both diverge to $+\infty$, both diverge to $-\infty$, or both converge in $\mathbb{R}$. Applying Lemma 1 with $\phi = \phi^{(l+1)}$, $f(t) = u_{sj}^{(l+1)}(t)$, and $g(t) = u_{sk}^{(l+1)}(t)$, we conclude that $\lim_{t\to+\infty} \Delta h_s^{(l+1)}(t)$ exists and is finite. This completes the induction, hence the limits exist for all neurons in layer $L-1$.

Step 3 (output layer). Since the output layer is linear,

$$\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)} = \sum_{s=1}^{n_{L-1}} W_s^{(L)} \, \Delta h_s^{(L-1)}(t). \tag{6.6}$$

Each summand has a finite limit by Step 2, hence the sum converges to a finite constant. Therefore $\lim_{t\to+\infty} \left[\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)}\right]$ exists and is finite, which is the coherence property. □

We now define four classes of activation functions, corresponding to the coherence types introduced in Subsection 6.3.1.

Definition 8 (Coherent class)**.** A function $\phi : \mathbb{R} \to \mathbb{R}$ belongs to the *coherent class* if the following two conditions hold:

(i) $\phi(v)$ is continuous for all $v \in \mathbb{R}$ and differentiable for all $v \neq 0$;

(ii) there exist constants $d_+^{(l)} \geq 0$, $d_-^{(l)} \geq 0$, $e_+^{(l)}, e_-^{(l)} \in \mathbb{R}$ such that

$$\lim_{v\to+\infty} \left[\phi^{(l)}(v) - (d_+^{(l)} v + e_+^{(l)})\right] = \lim_{v\to-\infty} \left[\phi^{(l)}(v) - (d_-^{(l)} v + e_-^{(l)})\right] = 0$$

with $\lim_{v\to+\infty} \phi^{(l)\prime}(v) = d_+^{(l)}$ and $\lim_{v\to-\infty} \phi^{(l)\prime}(v) = d_-^{(l)}$.

Definition 9 (Convergent class)**.** A function $\phi : \mathbb{R} \to \mathbb{R}$ belonging to the coherent class belongs to the *convergent class* if $d_+ = 0$ in Definition 8 (the function is saturating).

Definition 10 (Hypercoherent class)**.** A function $\phi : \mathbb{R} \to \mathbb{R}$ belonging to the coherent class belongs to the *hypercoherent class* if Definition 8 holds with $\phi(v) = d_+ v + e_+$ for all $v \geq v_0$ with some $v_0 \in \mathbb{R}$ and $d_+^{(l)} > 0$ (the function is linear and strictly increasing beyond a certain input threshold).

Definition 11 (Weakly coherent class)**.** A function $\phi : \mathbb{R} \to \mathbb{R}$ belonging to the coherent class belongs to the *weakly coherent class* if it belongs to neither the convergent nor the hypercoherent class, and $\exists v_0 \in \mathbb{R}$ such that $\phi'(v)$ is a strictly monotone function for all $v \geq v_0$.

Corollary 1. The mortality forecast generated by FeMoNet is coherent if all activation functions $\phi^{(l)}$ $(l = 1, \ldots, L-1)$ belong to the coherent class.

Proof of Corollary 1. It follows immediately from Theorem 1. □

In the sequel, we assume $\mathbf{w}^{(t)} > \mathbf{0}$ and $\mathbf{W}^{(l)} > \mathbf{0}$ $(\ell = 2, \ldots, L)$.[7] These constraints do not impair the representational capacity of the network, as population and age effects are modeled through embeddings whose components are unconstrained, thereby preserving flexibility. Moreover, it is reasonable to assume that the predictions $\hat{y}_{x,t}^{(p)} \triangleq -\log(m_{x,t}^{(p)})$ are increasing in t.

Theorem 2 (Closed form of asymptotic gap under positive downstream weights)**.** Consider the mortality forecast generated by FeMoNet, and assume $\mathbf{w}^{(t)} > \mathbf{0}$, $\mathbf{W}^{(l)} > \mathbf{0}$ $(l = 2, \ldots, L)$, and that the functions $\phi^{(l)}$ $(l = 1, \ldots, L-1)$ belong to the coherent class with constants $d_+^{(l)}, e_+^{(l)}$ as in Definition 8. Then for any $x_i \in \mathcal{X}$ and $j, k \in \mathcal{P}$,

$$\lim_{t\to+\infty} \left[\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)}\right] = \left(\prod_{\ell=1}^{L-1} d_+^{(\ell)}\right) \mathbf{W}^{(L)}\mathbf{W}^{(L-1)} \cdots \mathbf{W}^{(2)}\mathbf{W}^{(p)} \, (\mathbf{p}_j - \mathbf{p}_k). \tag{6.7}$$

[7] Positive weights can be achieved in practice by reparameterizing via $w = e^z$, where $z \in \mathbb{R}$, or by imposing the constraint $w \geq \varepsilon$ for some small $\varepsilon > 0$.

Lemma 2 (Asymptotic linearization)**.** Let $\phi : \mathbb{R} \to \mathbb{R}$ satisfy Conditions (i)–(ii) of Theorem 1. Let $f(t)$ and $g(t)$ be real-valued functions such that $\lim_{t\to+\infty}[f(t) - g(t)]$ exists and is finite. Assume moreover that either $f(t), g(t) \to +\infty$, or $f(t), g(t) \to -\infty$. Then

$$\lim_{t\to+\infty} \big[\phi(f(t)) - \phi(g(t))\big] = \begin{cases} d_+ \lim_{t\to+\infty}[f(t) - g(t)], & \text{if } f(t), g(t) \to +\infty, \\ d_- \lim_{t\to+\infty}[f(t) - g(t)], & \text{if } f(t), g(t) \to -\infty. \end{cases}$$

Proof of Lemma 2. Suppose $f(t), g(t) \to +\infty$ (the case $f(t), g(t) \to -\infty$ is analogous). For large enough t, the interval between $f(t)$ and $g(t)$ excludes 0. Hence by Condition (i) the Mean Value Theorem applies and yields a point $\xi(t)$ between $f(t)$ and $g(t)$ such that

$$\phi(f(t)) - \phi(g(t)) = \phi'(\xi(t))\,[f(t) - g(t)].$$

Since $\xi(t) \to +\infty$, Condition (ii) implies $\phi'(\xi(t)) \to d_+$. As $f(t)-g(t)$ has a finite limit, the product converges to $d_+ \lim_{t\to+\infty}[f(t) - g(t)]$, proving the claim. □

Proof of Theorem 2. Fix $x_i \in \mathcal{X}$ and $j, k \in \mathcal{P}$. For each layer $l = 1, \dots, L-1$ and neuron $s = 1, \dots, n_l$, define the differences

$$\Delta h_s^{(l)}(t) \triangleq h_s^{(l)}(j, t) - h_s^{(l)}(k, t), \qquad \Delta u_s^{(l)}(t) \triangleq u_{sj}^{(l)}(t) - u_{sk}^{(l)}(t).$$

Let $\Delta \mathbf{h}^{(l)}(t) \triangleq (\Delta h_1^{(l)}(t), \dots, \Delta h_{n_l}^{(l)}(t))^\top$.

Step 1 (base layer: $l = 1$). By (6.4), $\Delta u_s^{(1)}(t)$ is constant in t. Since $\mathbf{w}^{(t)} > \mathbf{0}$, we have $u_{sj}^{(1)}(t) \to +\infty$ and $u_{sk}^{(1)}(t) \to +\infty$. Applying Lemma 2 with $\phi = \phi^{(1)}$, $f(t) = u_{sj}^{(1)}(t)$, and $g(t) = u_{sk}^{(1)}(t)$ yields

$$\lim_{t\to+\infty} \Delta h_s^{(1)}(t) = d_+^{(1)}\, \Delta u_s^{(1)}(t) \qquad (s = 1, \dots, n_1).$$

Step 2 (induction). We prove by induction that for each $l = 1, \dots, L-1$,

$$\lim_{t\to+\infty} \Delta\mathbf{h}^{(l)}(t) = \left(\prod_{q=1}^{l} d_+^{(q)}\right) \mathbf{W}^{(l)}\mathbf{W}^{(l-1)} \cdots \mathbf{W}^{(2)}\mathbf{W}^{(p)} (\mathbf{p}_j - \mathbf{p}_k), \tag{6.8}$$

with the convention that the product of matrices is $\mathbf{W}^{(p)}$ when $l = 1$. The case $l = 1$ follows from Step 1 by stacking the components.

Assume (6.8) holds for some $l \in \{1, \dots, L-2\}$. Fix any $s = 1, \dots, n_{l+1}$. By (6.5),

$$\Delta u_s^{(l+1)}(t) = \sum_{r=1}^{n_l} W_{sr}^{(l+1)} \Delta h_r^{(l)}(t).$$

Taking $t \to +\infty$ and using the induction hypothesis yields that $\lim_{t\to+\infty} \Delta u_s^{(l+1)}(t)$ exists and equals the corresponding component of

$$\mathbf{W}^{(l+1)} \left(\prod_{q=1}^{l} d_+^{(q)}\right) \mathbf{W}^{(l)} \cdots \mathbf{W}^{(2)}\mathbf{W}^{(p)} (\mathbf{p}_j - \mathbf{p}_k).$$

If $\prod_{q=1}^{l} d_+^{(q)} = 0$, then $\lim_{t\to+\infty} \Delta u_s^{(l+1)}(t) = 0$. Since $u_{sj}^{(l+1)}(t)$ and $u_{sk}^{(l+1)}(t)$ share the same asymptotic behavior by Theorem 1, we obtain $\lim_{t\to+\infty} \Delta h_s^{(l+1)}(t) = 0$ by Lemma 2 (in the divergent cases) and the continuity of $\phi^{(l+1)}$ (in the convergent case), which agrees with (6.8) for layer $l+1$.

Otherwise $\prod_{q=1}^{l} d_+^{(q)} > 0$, hence $d_+^{(q)} > 0$ for all $q = 1, \dots, l$. Since $\mathbf{w}^{(t)} > \mathbf{0}$ and $\mathbf{W}^{(m)} > \mathbf{0}$ $(m = 2, \dots, l+1)$, induction over layers yields that $u_{sj}^{(l+1)}(t) \to +\infty$ and $u_{sk}^{(l+1)}(t) \to +\infty$. Applying Lemma 2 with $\phi = \phi^{(l+1)}$ gives

$$\lim_{t\to+\infty} \Delta h_s^{(l+1)}(t) = d_+^{(l+1)} \lim_{t\to+\infty} \Delta u_s^{(l+1)}(t),$$

which yields (6.8) for $l+1$. This completes the induction.

Step 3 (output layer). By (6.6),

$$\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)} = \sum_{s=1}^{n_{L-1}} W_s^{(L)} \Delta h_s^{(L-1)}(t).$$

Taking $t \to +\infty$ and using (6.8) with $l = L - 1$ yields (6.7). □

In the next theorem, we aim to provide sufficient conditions for the coherence subtypes defined in Subsection 6.3.1. In particular, weak coherence can be guaranteed under the following extra regularity conditions.[8] As embedding values remain unconstrained, these do not impair the representational capacity of the network.

Definition 12 (Regularity conditions for weak coherence)**.** FeMoNet satisfies the *regularity conditions for weak coherence* if

$$\exists j \neq k: \quad \mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \geq \mathbf{0}, \qquad \mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \neq \mathbf{0},$$

and the derivatives of all weakly coherent activation functions occurring in the network are eventually strictly monotone in the same direction.

Theorem 3 (Sufficient conditions for coherence classes)**.** Considering the mortality forecast generated by FeMoNet, if $\mathbf{w}^{(t)} > \mathbf{0}$, $\mathbf{W}^{(l)} > \mathbf{0}$ $(l = 2, \ldots, L)$, and $\phi^{(l)}$ $(l = 1, \ldots, L-1)$ belong to the coherent class then the following statements are true:

(i) the forecast is convergent if $\exists \ell \in \{1, \ldots, L-1\}$ such that $\phi^{(\ell)}$ belongs to the convergent class;

(ii) the forecast is hypercoherent if all $\phi^{(l)}$ $(l = 1, \ldots, L-1)$ belong to the hypercoherent class;

(iii) the forecast is weakly coherent if all $\phi^{(l)}$ $(l = 1, \ldots, L-1)$ belong to the weakly coherent class and the regularity conditions for weak coherence are satisfied;

(iv) the forecast is weakly coherent if $\nexists \ell$ such that $\phi^{(\ell)}$ belongs to the convergent class, $\exists \ell$ such that $\phi^{(\ell)}$ belongs to the weakly coherent class, and the regularity conditions for weak coherence are satisfied.

Proof of Theorem 3. Fix $x_i \in \mathcal{X}$ and $j, k \in \mathcal{P}$. Coherence follows from Corollary 1.

[8] These rule out the special cases when nonlinearities are canceled across neurons, leading to hypercoherent forecasts, and when the population embedding collapses all populations into the same value, resulting in convergent forecasts.

(i) If $\phi^{(\ell)}$ belongs to the convergent class then $d_+^{(\ell)} = 0$, and it follows from Theorem 2 that $\lim_{t\to+\infty}\left[\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)}\right] = 0$, proving convergence.

(ii) It suffices to show that $\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)}$ is constant for all large enough t. We have $u_{sj}^{(1)}(t) \to +\infty$ for all s due to $\mathbf{w}^{(t)} > \mathbf{0}$, and by (6.4) the differences $\Delta u_s^{(1)}(t)$ are constant in t. As $\phi^{(1)}$ belongs to the hypercoherent class, it is linear with slope $d_+^{(1)} > 0$ for large enough inputs, hence $\Delta h_s^{(1)}(t) = d_+^{(1)}\,\Delta u_s^{(1)}(t)$ for all large enough t, so $\Delta\mathbf{h}^{(1)}(t)$ is eventually constant. Assume $\Delta\mathbf{h}^{(l)}(t)$ is eventually constant for some $l \in \{1,\dots,L-2\}$. Then $\Delta\mathbf{u}^{(l+1)}(t) = \mathbf{W}^{(l+1)}\Delta\mathbf{h}^{(l)}(t)$ is eventually constant by (6.5), and we also have $u_{sj}^{(l+1)}(t) \to +\infty$ for all s due to $\mathbf{W}^{(l+1)} > \mathbf{0}$. Thus, $\phi^{(l+1)}$ is eventually linear, $\Delta\mathbf{h}^{(l+1)}(t)$ is eventually constant, and by induction, $\Delta\mathbf{h}^{(L-1)}(t)$ is eventually constant. Therefore $\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)} = \mathbf{W}^{(L)}\Delta\mathbf{h}^{(L-1)}(t)$ is eventually constant by (6.6), proving hypercoherence.

(iii) *Step 1 (not convergent).* Since $\phi^{(l)}$ is not in the convergent class, we have $d_+^{(l)} > 0$ for all l, hence $\prod_{l=1}^{L-1} d_+^{(l)} > 0$. By the regularity conditions, there exist $j \neq k$ such that

$$\mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \geq \mathbf{0}, \qquad \mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \neq \mathbf{0}.$$

Since $\mathbf{W}^{(l)} > \mathbf{0}$ $(l = 2,\dots,L)$,

$$\mathbf{W}^{(L)}\mathbf{W}^{(L-1)}\cdots\mathbf{W}^{(2)}\mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) > \mathbf{0}.$$

Therefore, Theorem 2 yields

$$\lim_{t\to+\infty}\left[\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)}\right] > 0,$$

so the forecast is not convergent.

Step 2 (not hypercoherent). Choose $j \neq k$ as above and suppose $\hat{y}_{x_i,t}^{(j)} - \hat{y}_{x_i,t}^{(k)}$ is constant for all $t \geq \tau$. Then

$$\partial_t\hat{y}_{x_i,t}^{(j)} = \partial_t\hat{y}_{x_i,t}^{(k)}$$

for all $t \geq \tau$.

Since

$$\Delta\mathbf{u}^{(1)}(t) = \mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \geq \mathbf{0}, \qquad \Delta\mathbf{u}^{(1)}(t) \neq \mathbf{0},$$

and $d_+^{(1)} > 0$, we have

$$\Delta \mathbf{h}^{(1)}(t) \geq \mathbf{0}, \qquad \Delta \mathbf{h}^{(1)}(t) \neq \mathbf{0}$$

for all sufficiently large t. Using $\mathbf{W}^{(2)} > \mathbf{0}$, it follows that

$$\Delta \mathbf{u}^{(2)}(t) > \mathbf{0}.$$

Inductively, using $\mathbf{W}^{(l)} > \mathbf{0}$ and $d_+^{(l)} > 0$, we obtain

$$u_{s,j}^{(l)}(t) > u_{s,k}^{(l)}(t)$$

for every hidden layer l, every neuron s, and all sufficiently large t. Moreover, $\mathbf{w}^{(t)} > \mathbf{0}$ implies

$$u_{s,j}^{(l)}(t), u_{s,k}^{(l)}(t) \to +\infty.$$

Define

$$\mathbf{G}^{(l)}(j,t) \triangleq \partial_t \mathbf{h}^{(l)}(j,t).$$

Then

$$\begin{aligned}
\mathbf{G}^{(1)}(j,t) &= \operatorname{diag}\big(\phi^{(1)\prime}(\mathbf{u}^{(1)}(j,t))\big)\mathbf{w}^{(t)}, \\
\mathbf{G}^{(l)}(j,t) &= \operatorname{diag}\big(\phi^{(l)\prime}(\mathbf{u}^{(l)}(j,t))\big)\mathbf{W}^{(l)}\mathbf{G}^{(l-1)}(j,t).
\end{aligned}$$

For sufficiently large t, all entries of $\mathbf{G}^{(l)}(\cdot,t)$ are strictly positive. By the regularity conditions, all weakly coherent activation derivatives are eventually strictly monotone in the same direction.

If these derivatives are eventually increasing,

$$\mathbf{G}^{(1)}(j,t) \geq \mathbf{G}^{(1)}(k,t), \qquad \mathbf{G}^{(1)}(j,t) \neq \mathbf{G}^{(1)}(k,t),$$

and positivity of the downstream weight matrices yields inductively

$$\mathbf{G}^{(l)}(j,t) > \mathbf{G}^{(l)}(k,t)$$

for every hidden layer l. If the derivatives are eventually decreasing, all inequalities

are reversed. Hence,

$$\mathbf{G}^{(L-1)}(j,t) \neq \mathbf{G}^{(L-1)}(k,t).$$

Since $\mathbf{W}^{(L)} > \mathbf{0}$,

$$\partial_t \hat{y}^{(j)}_{x_i,t} = \mathbf{W}^{(L)}\mathbf{G}^{(L-1)}(j,t) \neq \mathbf{W}^{(L)}\mathbf{G}^{(L-1)}(k,t) = \partial_t \hat{y}^{(k)}_{x_i,t},$$

contradicting constancy of the forecast difference. Hence the forecast is not hypercoherent.

Since it is coherent but neither convergent nor hypercoherent, it is weakly coherent.

(iv) Since $d_+^{(l)} > 0$ for all l, Theorem 2 and the regularity conditions imply that the forecast is not convergent.

Let r denote the first hidden layer whose activation belongs to the weakly coherent class. All preceding hidden layers, if any, belong to the hypercoherent class and therefore have constant positive tail derivatives. Consequently, the population ordering induced by

$$\mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \geq \mathbf{0}, \qquad \mathbf{W}^{(p)}(\mathbf{p}_j - \mathbf{p}_k) \neq \mathbf{0}$$

is preserved up to layer r. At layer r, strict monotonicity of $\phi^{(r)\prime}$ implies

$$\mathbf{G}^{(r)}(j,t) \neq \mathbf{G}^{(r)}(k,t).$$

Since all subsequent weakly coherent activation derivatives are eventually strictly monotone in the same direction, while all hypercoherent activation derivatives are constant, this ordering is preserved through the remaining hidden layers. Hence

$$\mathbf{G}^{(L-1)}(j,t) \neq \mathbf{G}^{(L-1)}(k,t),$$

and positivity of $\mathbf{W}^{(L)}$ gives

$$\partial_t \hat{y}^{(j)}_{x_i,t} \neq \partial_t \hat{y}^{(k)}_{x_i,t}.$$

Therefore, $\hat{y}^{(j)}_{x_i,t} - \hat{y}^{(k)}_{x_i,t}$ cannot be constant for all sufficiently large t, so the forecast

is not hypercoherent.

Therefore, the forecast is weakly coherent. □

Corollary 2. Let $\{\mathcal{P}_j\}_{j=1}^{q}$ be a partition of $\mathcal{P}$ as in Definition 6. For any coherence property $C \in \{\text{convergent}, \text{hypercoherent}, \text{weakly coherent}\}$, if the restriction of FeMoNet to each subgroup $\mathcal{P}_j$ satisfies the corresponding sufficient conditions of Theorem 3 for property C, then the resulting MMF is locally C with respect to the partition.

Proof of Corollary 2. By Theorem 3, the restriction of FeMoNet to each subgroup $\mathcal{P}_j$ has property C. Hence, by Definition 6, the MMF is locally C with respect to the partition. □

6.3.4 Activation functions

Table 6.1 lists several notable activation functions, their derivatives, and corresponding class memberships, which are straightforward to verify using elementary calculus.

Name	**Formula** $\phi(v) =$	**Derivative** $\phi'(v) =$	**Class**
Linear	v	1	hypercoherent
ReLU	$\max(0, v)$	$\begin{cases} 0, & v < 0 \\ 1, & v > 0 \end{cases}$	hypercoherent
Leaky ReLU[9] $(0 < \alpha < 1)$	$\begin{cases} \alpha v, & v < 0 \\ v, & v \geq 0 \end{cases}$	$\begin{cases} \alpha, & v < 0 \\ 1, & v > 0 \end{cases}$	hypercoherent
ELU[10] $(\alpha > 0)$	$\begin{cases} \alpha(e^v - 1), & v \leq 0 \\ v, & v > 0 \end{cases}$	$\begin{cases} \alpha e^v, & v < 0 \\ 1, & v > 0 \end{cases}$	hypercoherent
Sigmoid[11]	$\sigma(v) \triangleq \dfrac{1}{1 + e^{-v}}$	$\sigma(v)[1 - \sigma(v)]$	convergent
Tanh	$\tanh v$	$1 - \tanh^2 v$	convergent
Softsign	$\dfrac{v}{1 + \lvert v \rvert}$	$\dfrac{1}{(1 + \lvert v \rvert)^2}$	convergent
Swish[12]	$v\,\sigma(v)$	$\sigma(v)[1 + v(1 - \sigma(v)]$	weakly coherent
GELU[13]	$v\,\Phi(v)$	$\Phi(v) + v\,\varphi(v)$	weakly coherent
Softplus	$s(v) \triangleq \log(1 + e^v)$	$\sigma(v)$	weakly coherent
Mish	$v\,\tanh(s(v))$	$\tanh(s(v))$ $+v\,\operatorname{sech}^2(s(v))\,\sigma(v)$	weakly coherent

Table 6.1: Activation functions along with their derivatives and coherence classes

Additionally, we introduce the two-parameter families $\text{Swish}_{(\alpha,\beta)}$ $(\beta > 0)$ with

$$\phi(v) = v\,\sigma(\alpha + \beta v)$$
$$\phi'(v) = \frac{f(v)}{v}\{1 + \beta\,[v - f(v)]\},$$

[9] Called PReLU if α is a trainable parameter.
[10] The closely related SELU also belongs to the hypercoherent class.
[11] Also called Logistic.
[12] Also called SiLU.
[13] Φ and φ denote the cdf and pdf of the standard normal distribution.

and $\text{GELU}_{(\alpha,\beta)}$ $(\beta > 0)$ with

$$\phi(v) = v\,\Phi(\alpha + \beta v)$$
$$\phi'(v) = \frac{f(v)}{v} + \beta v\,\varphi(\alpha + \beta v),$$

both of which belong to the weakly coherent class. The parameter β allows these activations to interpolate smoothly between near-linear behavior ($\beta \to 0$) and ReLU-like sharpness ($\beta \to \infty$), while α provides additional flexibility. The availability of two free parameters may facilitate the calibration of the model to simultaneously achieve good predictive performance and empirically plausible coherence behavior.

Richman and Wüthrich (2021) employed ReLU and Tanh activation functions, which, by Theorem 3, result in hypercoherent and convergent forecasts, respectively. While their work represents an important step towards the application of neural networks for mortality forecasting, from a practical perspective, both behaviors may be unnecessarily restrictive. Hypercoherence implies that log-mortality forecasts become parallel beyond some time threshold, whereas in practice differences between populations often continue to evolve over time, and the onset of parallelism cannot be controlled independently of the learned network parameters. Convergence is even more restrictive, as it builds asymptotic equality of mortality forecasts across populations directly into the network architecture. These observations motivate the use of activation functions from the weakly coherent class.

6.4 Empirical results

6.4.1 Data and sample construction

The empirical analysis uses annual deaths and exposures from the Human Mortality Database (HMD, 2025), observed by single year of age and separately for men and women. The sample covers 26 countries: Australia, Austria, Belgium, Canada, Czechia, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Japan, Latvia, Lithuania, the Netherlands, Norway, Poland, Portugal, Slovakia,

Spain, Sweden, Switzerland, the United Kingdom, and the United States. Male and female observations are treated as distinct populations. The resulting panel therefore contains 52 country–sex populations.

For every population, ages 0–99 are retained. The estimation period is 1960–2009, while 2010–2019 is reserved as a genuinely out-of-time forecast sample. Thus, an observation is identified by a country–sex population, a calendar year, and an age. The estimation panel contains $52 \times 50 \times 100 = 260{,}000$ observations, and the final evaluation panel contains $52 \times 10 \times 100 = 52{,}000$ observations.

The central mortality rate is calculated as the ratio of deaths to exposure. To avoid undefined logarithms in cells with zero or very few deaths, a Haldane–Anscombe-type correction of 0.5 is added to each death count. The modelled quantity is therefore

$$\log m_{x,t}^{(p)} = \log\left(\frac{D_{x,t}^{(p)} + 0.5}{E_{x,t}^{(p)}}\right), \tag{6.9}$$

where D denotes deaths, E denotes exposure, x is age, t is calendar year, and p indexes the country–sex population. The neural networks are trained on $-\log m$; their forecasts are transformed back to the log-mortality scale before evaluation. Calendar year is linearly scaled to the $[0, 1]$ interval using the endpoints of the 1960–2009 estimation period.

6.4.2 Benchmark models

The conventional benchmark is the Li–Lee multipopulation mortality model, estimated separately for all 52 country–sex populations while preserving a shared mortality component. Estimation is performed in R with the `fit_li_lee` routine of the `MultiMoMo` package, using Newton–Raphson estimation on the 1960–2009 period.

The common time index is forecast as a random walk with drift. Two alternative specifications are used for the population-specific time indices. In the first, each specific index follows a random walk without drift (Li–Lee RW). In the second, it follows an AR(1) process, estimated as an ARIMA$(1, 0, 0)$ model by conditional sum of squares (Li–Lee AR(1)). Both specifications produce ten-

year point forecasts for 2010–2019. These alternatives make it possible to assess whether short-run mean reversion in population-specific deviations improves on a persistent random-walk specification.

6.4.3 Neural-network specifications

Two constrained feed-forward neural networks are compared with the Li–Lee benchmarks. Both receive three inputs: scaled calendar year, age, and the country–sex population identifier. Age and population are passed through separate five-dimensional embedding layers.

The weights of the dense layers are subject to a positivity constraint. The networks have five hidden layers with 256, 128, 128, 64 and 32 units. In both cases, the final layer contains one linear output unit with the same positive-weight parameterisation.

Both networks are trained by the Adam optimiser with a learning rate of 0.001, a minibatch size of 32, mean squared training loss, and gradient-norm clipping at one.

To reduce sensitivity to random initialisation, both neural specification is estimated with ten random seeds. The prediction used in the comparison is the arithmetic mean of these ten forecasts on the log-mortality scale. Consequently, the reported ReLU and Swish results refer to neural-network ensembles rather than to a single fitted network. A Lee–Miller jump-off correction (Lee and Miller, 2001) anchored at the final observed year, 2009, is applied before the forecasts are compared, ensuring that all methods start from a common observed endpoint.

6.4.4 Evaluation and empirical findings

The final comparison is based on country-level forecast performance over the common 2010–2019 out-of-time panel. Rather than relying on a single pooled summary statistic, the evaluation asks how far each model falls behind the best-performing model within each country–sex population. For model m and population p, absolute underperformance is defined as

$$UP_m^{(p)} = L_m^{(p)} - \min_j L_j^{(p)}, \tag{6.10}$$

where $L_{m,p}$ is the population-level forecast loss and the minimum is taken over the four competing models. A value of zero means that the model is best for that population; larger values indicate a greater distance from the local winner. Figure 6.4 reports the median, mean, and maximum of this distance across the 52 populations, together with the number of populations won by each model. Lower bars are preferable in the first three panels, whereas a higher bar is preferable in the final panel.

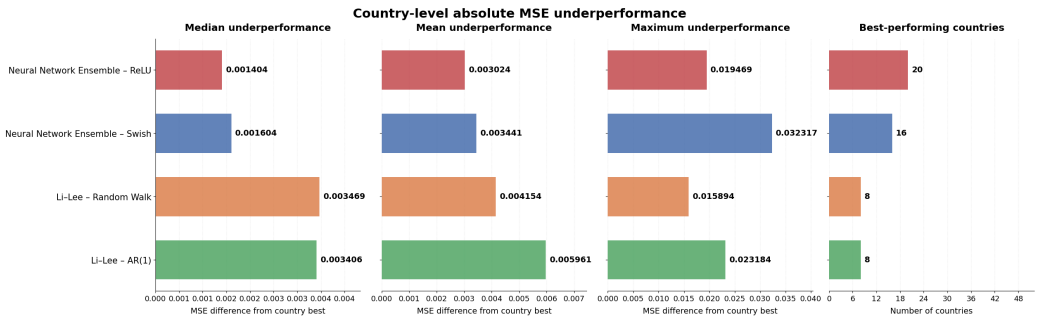

Figure 6.4: Country-level absolute forecast underperformance of the ReLU and Swish neural-network ensembles and the two Li–Lee benchmarks, 2010–2019.

The figure shows that the ReLU ensemble provides the strongest overall country-level performance. It has the lowest median and mean underperformance, indicating that it typically remains closest to the best model in each population. It is also the best-performing specification in 20 of the 52 country–sex populations, the highest count among the four methods. The Swish ensemble ranks second on the median and mean criteria and wins 16 populations. Taken together, the two neural approaches are therefore best in 36 of the 52 populations, or slightly more than two thirds of the panel.

The Li–Lee random-walk and AR(1) specifications each win eight populations. Their typical distance from the population-specific winner is larger than that of either neural ensemble. The comparison of the two Li–Lee variants is not uniform across the panels: the AR(1) version has a slightly lower median underperformance, but the random-walk version has a lower mean and a lower worst-case

shortfall. This suggests that mean reversion in the specific time indices helps in some populations but does not provide a consistently more robust forecast than the simpler random-walk continuation.

The maximum-underperformance panel also qualifies the neural-network result. Although the Swish ensemble performs well on average, it exhibits the largest worst-case shortfall of the four models. The ReLU ensemble has a substantially smaller maximum shortfall than Swish and Li–Lee AR(1), although Li–Lee RW is slightly more favourable on this single worst-case criterion. Hence, the empirical advantage of the ReLU ensemble is not the consequence of dominating every country–sex population. Instead, it combines the broadest number of population-level wins with the smallest typical loss relative to the local best method.

Overall, the evidence in Figure 6.4 supports a measured conclusion. A positive-weight neural network with age and population embeddings can improve coherent multipopulation mortality forecasting relative to the Li–Lee benchmarks, particularly when ReLU activations and an ensemble of initialisations are used. Despite its somewhat weaker empirical performance, the Swish specification has the theoretical advantage of producing weakly coherent forecasts.

6.5 Conclusion

This chapter developed a theoretical framework for coherent multipopulation mortality forecasting using deep feedforward neural networks. We established sufficient conditions under which feedforward networks generate coherent forecasts, introduced a taxonomy distinguishing convergent, hypercoherent, weakly coherent and incoherent forecasts, and identified corresponding classes of activation functions. We further derived explicit sufficient conditions under which these coherence subclasses arise, thereby connecting architectural choices to long-run forecasting behaviour.

Several limitations should nevertheless be acknowledged. First, some assumptions, most notably positivity of downstream weights, are primarily introduced to obtain transparent mathematical results and may not be required in practice. Second, the theoretical analysis focuses on point forecasts and does not address predictive uncertainty. Third, the empirical study is intentionally limited in scope

and serves primarily as an illustration of the theoretical framework rather than a comprehensive benchmarking exercise.

The proposed framework suggests several directions for future research. An important extension would be to develop coherent prediction intervals and probabilistic forecasts by combining the present theory with uncertainty quantification techniques. More extensive empirical validation across additional countries would help assess the practical robustness of the proposed methodology. The theory could also be extended to architectures imposing only local coherence, thereby accommodating heterogeneous groups of populations. Another natural direction is the incorporation of additional actuarial shape constraints, such as age-monotonicity or smoothness of mortality curves, directly into the network architecture or optimization procedure. It would furthermore be interesting to investigate whether analogous coherence guarantees can be established for more recent neural architectures, including recurrent networks, transformers, Kolmogorov–Arnold networks, or graph neural networks.

Overall, the results demonstrate that coherence can be viewed as an architectural property of deep neural networks, combining theoretical consistency with the flexibility required to capture complex mortality dynamics. The proposed framework therefore provides a practical foundation for constructing coherent neural mortality forecasting models suitable for practical actuarial applications.

Summary and Conclusions

In the following, I refer back to the research questions outlined in Section 1.3 and briefly summarize the answers that can be provided to them based on the findings of the studies. This chapter does not repeat detailed study-by-study summaries of the three studies, as their main conclusions, limitations, and directions for future research are presented at the end of the corresponding chapters.

- Research question 1: Does using more complex time series methods to forecast the Lee–Carter model's mortality index lead to improved forecast accuracy? Which method provides the most reliable forecasts for actuaries and demographers?

- Answer 1: The analysis of European countries demonstrates that more advanced time series and econometric methods have clear practical value. Actuaries and demographers can improve the accuracy of their forecasts by moving beyond traditional Random Walk or ARIMA models and adopting alternative approaches when forecasting the mortality index in Lee–Carter or Li–Lee frameworks. In many cases, the most accurate forecasts were produced by methods that account for spatial relationships. This allows geographic dependencies to be incorporated into the forecasting framework, enhancing overall model performance.

- Research question 2: To what extent can forecasts of mortality rates in the Visegrád Group countries be improved by using neural networks compared to classical methods? Do neural networks provide more precise forecasts across all populations of the Visegrád Group countries? Are the forecast results significantly better?

- Answer 2: When considering aggregate performance metrics, various recurrent neural network models perform very well compared to the classical Lee–Carter and Li–Lee approaches. However, a more granular analysis by age group reveals a different pattern: the Li–Lee model tends to provide the most accurate forecasts for older age groups, whereas neural network–based methods achieve superior performance primarily for younger and middle-aged populations.

- Research question 3: Can neural networks produce coherent mortality forecasts? If so, under what conditions?

- Answer 3: Yes, it can be formally shown that, in the case of feedforward neural networks, coherent forecasts can be obtained for certain activation functions when the architecture described in the study is applied.

References

Allaire, J. J. & Chollet, F. (2022). keras: R interface to Keras https://CRAN.R-project.org/package=keras.

Álvaro-Meca, A. et al. (2011). The Convergence of European Mortality in Both Sexes in the Near Future: A Spatio-Temporal Approach. *Population Review*, *50*(2), https://doi.org/10.1353/prv.2011.0018.

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115, https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Arató, M., Bozsó, D., Elek, P., & Zempléni, A. (2009). Forecasting and simulating mortality tables. *Mathematical and Computer Modelling*, *49*(3-4), 805–813.

Arellano, M. & Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, *58*(2), 277–297, https://doi.org/10.2307/2297968.

Astuti, A. M., Setiawan, Zain, I., & Purnomo, J. D. T. (2020). A Review of Panel Data on Spatial Econometrics Models. *Journal of Physics: Conference Series*, *1490*(1), 012032, https://doi.org/10.1088/1742-6596/1490/1/012032.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Bennett, J. E., et al. (2015). The future of life expectancy and life expectancy inequalities in England and Wales: Bayesian spatiotemporal forecasting. *Lancet*, *386*(9989), 163–170, https://doi.org/10.1016/S0140-6736(15)60296-3.

Blundell, R. & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115–143, https://doi.org/10.1016/S0304-4076(98)00009-8.

Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee–Carter under conditions of variable mortality decline. *Population Studies*, *56*(3), 325–336, https://doi.org/10.1080/00324720215935.

Bravo, J. M. (2021). Forecasting mortality rates with recurrent neural networks: A preliminary investigation using portuguese data. In *CAPSI 2021 Proceedings*, number 7 https://aisel.aisnet.org/capsi2021/7.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Brouhns, N., Denuit, M., & Keilegom, I. V. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, *2005*(3), 212–224, https://doi.org/10.1080/03461230510009754.

Brouhns, N., Denuit, M., & Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, *31*(3), 373–393, https://doi.org/10.1016/S0167-6687(02)00185-3.

Cairns, A. J. (2000). A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, *27*(3), 313–330, https://doi.org/10.1016/S0167-6687(00)00055-X.

Cairns, A. J. G., Blake, D., & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: The-

ory and calibration. *Journal of Risk and Insurance*, *73*(4), 687–718, https://doi.org/10.1111/j.1539-6975.2006.00195.x http://dx.doi.org/10.1111/j.1539-6975.2006.00195.x.

Carracedo, P., Debón, A., Iftimi, A., et al. (2018). Detecting spatio-temporal mortality clusters of European countries by sex and age. *International Journal for Equity in Health*, *17*(1), 38, https://doi.org/10.1186/s12939-018-0750-z.

Chen, Y. & Khaliq, A. Q. M. (2022). Comparative study of mortality rate prediction using data-driven recurrent neural networks and the Lee–Carter model. *Big Data and Cognitive Computing*, *6*(4), 134, https://doi.org/10.3390/bdcc6040134.

Cheysson, F. (2016). *starma: Modelling Space Time AutoRegressive Moving Average (STARMA) Processes*. R package version 1.3, URL: https://CRAN.R-project.org/package=starma.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cliff, A. D. & Ord, J. K. (1973). *Spatial Autocorrelation*. London: Pion.

Costantini, M. & Lupi, C. (2013). A Simple Panel-CADF Test for Unit Roots. *Oxford Bulletin of Economics and Statistics*, *75*(2), 276–296, https://doi.org/10.1111/j.1468-0084.2012.00690.x.

Croissant, Y. & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43, https://doi.org/10.18637/jss.v027.i02.

Currie, I. D., Durban, M., & Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, *4*(4), 279–298, https://doi.org/10.1191/1471082X04st080oa.

de Moivre, A. (1752). *Annuities on Lives: With Several Tables, Exhibiting at One View, the Values of Lives, for Different Rates of Interest*. Oxford: A. Millar, 4th edition https://books.google.hu/books/about/Annuities_on_Lives.html?id=id5bAAAAQAAJ.

De Mori, L., Haberman, S., Millossovich, P., & Zhu, R. (2025). Mortality forecasting via multi-task neural networks. *ASTIN Bulletin*, *55*(2), 313–331, https://doi.org/10.1017/asb.2025.10.

Derksen, S. & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282, https://doi.org/10.1111/j.2044-8317.1992.tb00992.x.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26, https://doi.org/10.1214/aos/1176344552 https://doi.org/10.1214/aos/1176344552.

EIOPA (2020). *Discussion Paper on (Re)insurance Value Chain and Artificial Intelligence*. Technical report, European Insurance and Occupational Pensions Authority, Frankfurt am Main, Germany. Accessed: 10-26-2025 https://www.eiopa.europa.eu/system/files/2020-06/discussion-paper-on-insurance-value-chain-and-new-business-models-arising-fr pdf.

Elhorst, J. (2014). *Spatial econometrics: from cross-sectional data to spatial panels*. Springer. ISBN 9783642403408. https://link.springer.com/book/10.1007/978-3-642-40340-8.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Enchev, V., Kleinow, T., & Cairns, A. J. G. (2017). Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, *2017*(4), 319–342.

Euthum, M., Scherer, M., & Ungolo, F. (2024). A neural network approach for the mortality analysis of multiple populations: a case study on data of

the Italian population. *European Actuarial Journal*, *14*, 495–524, https://doi.org/10.1007/s13385-024-00377-5.

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of ICML*, (pp. 148–156).

Friedberg, S., Insel, A. J., & Spence, L. E. (2002). *Linear Algebra*. Pearson, 4th edition.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gibbs, Z., Groendyke, C., Hartman, B., & Richardson, R. (2020). Modeling county-level spatio-temporal mortality rates using dynamic linear models. *Risks*, *8*(4), https://doi.org/10.3390/risks8040117.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378, https://doi.org/10.1198/016214506000001437.

Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. f. r. s. &c. *Philosophical Transactions of the Royal Society of London*, *115*, 513–583, https://doi.org/10.1098/rstl.1825.0026.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, *74*, 537–552.

Greco, F. & Scalone, F. (2013). A space-time extension of the Lee–Carter model in a hierarchical Bayesian framework: Modelling and forecasting provincial mortality in Italy https://api.semanticscholar.org/CorpusID:55707767.

Griffith, D. A. (2010). Modeling spatio-temporal relationships: retrospect and prospect. *Journal of Geographical Systems*, *12*(2), 111–123, https://doi.org/10.1007/s10109-010-0120-x.

Griffith, D. A. (2012). Space, time, and space-time eigenvector filter specifications that account for autocorrelation. *Estadistica Española*, *54*, 4–27 https://www.researchgate.net/publication/288942431_Space_time_and_space-time_eigenvector_filter_specifications_that_account_for_autocorrelation.

Griffith, D. A. & Paelinck, J. H. P. (2018). *Morphisms for Quantitative Spatial Analysis*. Cham, Switzerland: Springer. Chapter "Space–Time Autocorrelation", pp. 25–34.

Guibert, Q., Lopez, O., & Piette, P. (2019). Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance: Mathematics and Economics*, *88*, 255–272, https://doi.org/10.1016/j.insmatheco.2019.07.004.

Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin*, *48*(2), 481–508, https://doi.org/10.1017/asb.2017.33.

Hassan, A. (2021). *Spatial data analysis: applications to population health*. PhD thesis, Université de Lille https://www.researchgate.net/publication/361052863_Spatial_data_analysis_applications_to_population_health.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2 edition.

HMD (2025). Human Mortality Database. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). http://www.mortality.org. Accessed: 21-06-2025.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

Hofstede, G. (1984). Cultural dimensions in management and planning. *Asia Pacific Journal of Management*, *1*(2), 81–99, https://doi.org/10.1007/BF01733682.

Hurlin, C. & Mignon, V. (2007). *Second Generation Panel Unit Root Tests*. HAL Working Papers halshs-00159842 https://ideas.repec.org/p/hal/wpaper/halshs-00159842.html.

Hyndman, R. J., Booth, H., & Yasmeen, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography*, *50*(1), 261–283, https://doi.org/10.1007/s13524-012-0145-5 https://doi.org/10.1007/s13524-012-0145-5.

Hyndman, R. J. & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1–22, https://doi.org/10.18637/jss.v027.i03.

Hyndman, R. J. & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, *51*(10), 4942–4956, https://doi.org/10.1016/j.csda.2006.07.028.

Islam, M. D., Li, B., Lee, C., & Wang, X. (2022). Incorporating spatial information in machine learning: The Moran eigenvector spatial filter approach. *Transactions in GIS*, *26*, 1–21, https://doi.org/10.1111/tgis.12894.

Jarner, S. F. & Jallbjørn, S. (2020). Pitfalls and merits of cointegration-based mortality models. *Insurance: Mathematics and Economics*, *90*, 80–93.

Jordan, M. I. (1986). Serial order: A parallel distributed processing approach. *Advances in Psychology*, *121*, 471–495.

Józan, P. (2002). A halandóság alapirányzata a 20. században, és az ezredforduló halálozási viszonyai magyarországon. *Magyar Tudomány*, (4), 419–439.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleiber, C. & Lupi, C. (2011). Panel Unit Root Testing with R https://api.semanticscholar.org/CorpusID:197629737.

Kleinov, T. (2015). A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, *63*, 147–152, https://doi.org/10.1016/j.insmatheco.2015.03.023 https://doi.org/10.1016/j.insmatheco.2015.03.023.

Koissi, M. C., Shapiro, A. F., & Hognas, G. (2006). Evaluating and extending the lee–carter model for mortality forecasting confidence interval. *Insurance: Mathematics and Economics*, *38*(1), 1–20, https://doi.org/10.1016/j.insmatheco.2005.06.008 https://doi.org/10.1016/j.insmatheco.2005.06.008.

Krúmiņš, J. & Zvidriņš, P. (1992). Recent mortality trends in the three baltic republics. *Population Studies*, *46*(2), 259–273, https://doi.org/10.1080/0032472031000146226.

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.

Lee, R. D. & Miller, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography*, *38*(4), 537–549, https://doi.org/10.1353/dem.2001.0036.

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.

Lindholm, M. & Palmborg, L. (2022). Efficient use of data for LSTM mortality forecasting. *European Actuarial Journal*, *12*, 749–778, https://doi.org/10.1007/s13385-022-00307-3.

Liu, Z. (2021). *Bayesian Poisson Log-normal Model with Regularized Time Structure and Spatial Framework for Mortality Projection of Multi-population*.

PhD dissertation, Clemson University https://tigerprints.clemson.edu/all_dissertations/2859.

Makeham, W. (1867). On the law of mortality. *Journal of the Institute of Actuaries*, *13*(6), 325–358 http://www.jstor.org/stable/41134517.

Meslé, F. & Vallin, J. (2002). Mortality in europe: the divergence between east and west. *Population (English edition)*, *57*(1), 157–197, https://doi.org/10.2307/3246630. https://hal.science/hal-04275601.

Miyata, A. & Matsuyama, N. (2022). Extending the Lee–Carter model with variational autoencoder: a fusion of neural network and Bayesian approach. *ASTIN Bulletin*, *52*(3), 789–812, https://doi.org/10.1017/asb.2022.15.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1–2), 17–23, https://doi.org/10.2307/2332142.

Májer, I. & Kovács, E. (2011). Élettartam-kockázat - a nyugdíjrendszerre nehezedő egyik teher. *Statisztikai Szemle*, *89*(7-8), 790–812.

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., & Perla, F. (2019). A deep learning integrated Lee–Carter model. *Risks*, *7*(1), 33, https://doi.org/10.3390/risks7010033.

Ocaña-Riola, R. & Mayoral-Cortés, J. M. (2010). Spatio-temporal trends of mortality in small areas of Southern Spain. *BMC Public Health*, *10*, 26, https://doi.org/10.1186/1471-2458-10-26.

Omran, A. R. (1971). The epidemiological transition. *Milbank Memorial Fund Quarterly*, *49*(4), 509–538, https://doi.org/10.2307/3349375.

Perla, F., Richman, R., Scognamiglio, S., & Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, *2021*(7), 572–598, https://doi.org/10.1080/03461238.2020.1867232.

Perla, F., Richman, R., Scognamiglio, S., & Wüthrich, M. V. (2024). Accurate and explainable mortality forecasting with the LocalGLMnet. *Scandinavian Actuarial Journal*, *2024*(7), 739–761, https://doi.org/10.1080/03461238.2024.2307620.

Perla, F. & Scognamiglio, S. (2023). Locally-coherent multi-population mortality modelling via neural networks. *Decisions in Economics and Finance*, *46*, 157–176, https://doi.org/10.1007/s10203-022-00382-x.

Petneházi, G. & Gáll, J. (2019). Mortality rate forecasting: can recurrent neural networks beat the lee–carter model? arXiv preprint arXiv:1909.05501 https://doi.org/10.48550/arXiv.1909.05501.

Pfeifer, P. E. & Deutsch, S. J. (1980). A three-stage iterative procedure for space-time modeling. *Technometrics*, *22*(1), 35–47 http://www.jstor.org/stable/1268381.

Pickle, L. W. (2000). Exploring spatio-temporal patterns of mortality using mixed effects models. *Statistics in Medicine*, *19*(17-18), 2251–2263, https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2251::AID-SIM567>3.0.CO;2-M.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, *45*(3), 393–404, https://doi.org/10.1016/j.insmatheco.2009.08.006 http://dx.doi.org/10.1016/j.insmatheco.2009.08.006.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Quinlan, J. R. (1996). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

R Core Team (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Renshaw, A. E. & Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, *38*(3), 556–570, https://doi.org/10.1016/j.insmatheco.2005.12.001.

Richman, R. (2021). AI in actuarial science: a review of recent advances – part 2. *Annals of Actuarial Science*, *15*(2), 230–258, https://doi.org/10.1017/S174849952000024X.

Richman, R. (2022). Mind the gap: safely incorporating deep learning models into the actuarial toolkit. *British Actuarial Journal*, *27*, e21, https://doi.org/10.1017/S1357321722000162.

Richman, R. & Wüthrich, M. V. (2019). A neural network extension of the lee-carter model to multiple populations. *Annals of Actuarial Science*, *13*(2), 370–389.

Richman, R. & Wüthrich, M. V. (2021). A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, *15*(2), 346–366, https://doi.org/10.1017/S1748499519000071.

Richman, R. & Wüthrich, M. V. (2022). LocalGLMnet: interpretable deep learning for tabular data. *Scandinavian Actuarial Journal*, *2023*(1), 71–95, https://doi.org/10.1080/03461238.2022.2081816.

Robben, J. (2021). *MultiMoMo: Multi-population mortality models*. R package version 1.0.0 https://github.com/jensrobben/MultiMoMo.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Roshani, A., Izadi, M., & Khaledi, B. (2022). Transformer self-attention network for forecasting mortality rates. *Journal of the Iranian Statistical Society*, *21*(1), 81–103, https://doi.org/10.22034/jirss.2022.704621.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536, https://doi.org/10.1038/323533a0 https://doi.org/10.1038/323533a0.

Salima, B. A. (2018). Spatial econometrics on panel data https://api.semanticscholar.org/CorpusID:214675854.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

Schnürch, S. & Korn, R. (2022). Point and interval forecasts of death rates using neural networks. *ASTIN Bulletin*, *52*(1), 333–360, https://doi.org/10.1017/asb.2021.34.

Scognamiglio, S. (2022). Calibrating the Lee–Carter and the Poisson Lee–Carter models via neural networks. *ASTIN Bulletin*, *52*(2), 519–561, https://doi.org/10.1017/asb.2022.5.

Simonovska, R. (2025). SDPDmod: Spatial Dynamic Panel Data Modeling https://CRAN.R-project.org/package=SDPDmod.

Szentkereszti, G. & Vékás, P. (2026). Mortality forecasting in geographical space and time. *ASTIN Bulletin: Journal of the International Actuarial Association*.

Szentkereszti, G. & Vékás, P. (2022). Magyar halandósági ráták előrejelzése visszacsatolt neurális hálózatokkal. *Statisztikai Szemle*, *100*(10), 905–922, https://doi.org/10.20311/stat2022.10.hu0905 https://doi.org/10.20311/stat2022.10.hu0905.

Szentkereszti, G. & Vékás, P. (2024). A visegrádi országok mortalitási rátáinak előrejelzése neurális hálózatokkal. *Szigma*, *55*(2–3), 293–312, https://doi.org/10.15170/SZIGMA.55.1242 https://doi.org/10.15170/SZIGMA.55.1242.

Tanaka, S. & Matsuyama, N. (2025). An interpretable neural network approach to cause-of-death mortality forecasting. *Annals of Actuarial Science*, (pp. 1–20)., https://doi.org/10.1017/S1748499524000319.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*(Supplement), 234–240, https://doi.org/10.2307/143141.

Tuljapurkar, S., Li, N., & Boe, C. (2000). A universal pattern of mortality decline in the G7 countries. *Nature*, *405*(6788), 789–792, https://doi.org/10.1038/35015561.

Tóth, C. D. (2021). Multi-population models to handle mortality crises in forecasting mortality: A case study from hungary. *Society and Economy*, *43*(2), 128–146, https://doi.org/10.1556/204.2021.00007 https://doi.org/10.1556/204.2021.00007.

Ugarte, M. D., Goicoa, T., & Militino, A. F. (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics*, *21*(3-4), 270–289, https://doi.org/10.1002/env.1011.

Vallin, J. & Meslé, F. (2004). Convergences and divergences in mortality. a new approach to health transition. *Demographic Research, Special Collection*, *2*(2), 1–43.

Van Belle, G. (2008). *Statistical Rules of Thumb*. Hoboken, NJ: Wiley, 2nd edition.

Vaswani, A., et al. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 5998–6008).

Vazzoler, S. (2021). *sparsevar: Sparse VAR/VECM Models Estimation*. R package version 0.1.0, URL: https://CRAN.R-project.org/package=sparsevar.

Villegas, A. M., Kaishev, V. K., & Millossovich, P. (2015). StMoMo: An R package for stochastic mortality modeling. *European Actuarial Journal*, *5*(2), 271–284, https://doi.org/10.18637/jss.v084.i03.

Vékás, P. (2017). Nyugdíjcélú életjáradékok élettartam-kockázata az általánosított korcsoport-időszak-kohorsz modellkeretben. *Statisztikai Szemle*, *95*(2), 139–165, https://doi.org/10.20311/stat2017.02.hu0139 https://doi.org/10.20311/stat2017.02.hu0139.

Vékás, P. (2019a). *Az élettartam-kockázat modellezése*. Budapest: Budapesti Corvinus Egyetem.

Vékás, P. (2019b). Rotation of the age pattern of mortality improvements in the European Union. *Central European Journal of Operations Research*, *28*(3), 1031–1048, https://doi.org/10.1007/s10100-019-00617-0.

Wang, C.-W., Zhang, J., & Zhu, W. (2021). Neighbouring prediction for mortality. *ASTIN Bulletin*, *51*(3), 689–718, https://doi.org/10.1017/asb.2021.13.

Wang, J., Wen, L., Xiao, L., & Wang, C. (2023). Time-series forecasting of mortality rates using transformer. *Scandinavian Actuarial Journal*, *2024*(2), 109–123, https://doi.org/10.1080/03461238.2023.2218859.

Wen, J., Cairns, A. J. G., & Kleinow, T. (2020). Fitting multi-population mortality models to socio-economic groups. *Annals of Actuarial Science*, *14*(1), 180–204, https://doi.org/10.1017/S1748499520000184 https://doi.org/10.1017/S1748499520000184.

Wilmoth, J. (1993). *Computational Methods for Fitting and Extrapolating the Lee–Carter Model of Mortality Change*. Technical report, Department of Demography, University of California, Berkeley http://demog.berkeley.edu/~jrw/Papers/LCtech.pdf.

Wüthrich, M. V. & Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications*. Springer Actuarial. Springer Cham https://link.springer.com/book/10.1007/978-3-031-12409-9.

Wüthrich, M. V., et al. (2025). *AI Tools for Actuaries*. SSRN eLibrary https://ssrn.com/abstract=5162304.

Wüthrich, M. V. & Merz, M. (2022). *Statistical Foundations of Actuarial Learning and Its Applications*. Springer.

Yu, J., de Jong, R., & Lee, L. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics*, *146*(1), 118–134, https://doi.org/10.1016/j.jeconom.2008.08.002.

Yu, J. & Lee, L. (2010). Estimation of unit root spatial dynamic panel data models. *Econometric Theory*, *26*(5), 1332–1362 http://www.jstor.org/stable/40800885.

Zhang, L. & Zhuang, Y. (2026). What KAN mortality say: smooth and interpretable mortality modeling using Kolmogorov–Arnold networks. *ASTIN Bulletin*, *56*(1), 32–59, https://doi.org/10.1017/asb.2025.10079.

Zheng, H., Wang, H., Zhu, R., & Xue, J.-H. (2025). A brief review of deep learning methods in mortality forecasting. *Annals of Actuarial Science*, *online first*, 1–16, https://doi.org/10.1017/S1748499525100110.

Őri, P. & Spéder, Z. (2020). Folytonos átmenet: Magyarország népesedése 1920 és 2020 között. *Statisztikai Szemle*, *98*(6), 481–521, https://doi.org/10.20311/stat2020.6.hu0481 https://doi.org/10.20311/stat2020.6.hu0481.

Publications of the Author

- Szentkereszti, G. & Vékás, P. (2022). Magyar halandósági ráták előrejelzése visszacsatolt neurális hálózatokkal. *Statisztikai Szemle*, 100(10), 905–922. https://doi.org/10.20311/stat2022.10.hu0905
- Szentkereszti, G. & Vékás, P. (2024). A visegrádi országok mortalitási rátáinak előrejelzése neurális hálózatokkal. *Szigma*, 55(2–3), 293–312. https://doi.org/10.15170/SZIGMA.55.1242
- Szentkereszti, G. & Vékás, P. (2026). Mortality forecasting in geographical space and time. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 56(2), 367–388. https://doi.org/10.1017/asb.2026.10092