Thesis submitted in fulfillment of the requirements for the degree

Dr. rer. pol.

on the topic

Advances in Bayesian Demographic Forecasting

to the Institute of Statistics

Faculty of Social Sciences, Economics, and Business Administration

University of Bamberg

submitted by

Julius Goes

Bamberg 2026

Cumulative dissertation
Julius Goes, *Advances in Bayesian Demographic Forecasting*

This thesis has been submitted to the Faculty of Social Sciences, Economics and Business Administration at the University of Bamberg as a dissertation.

First supervisor: Prof. Dr. Anne Leucht
Second supervisor: Prof. Dr. Henriette Engelhardt-Wölfler
Date of defense: 21.11.2025

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.

Erstgutachterin: Prof. Dr. Anne Leucht
Zweitgutachterin: Prof. Dr. Henriette Engelhardt-Wölfler
Tag der mündlichen Prüfung: 21.11.2025

URN: urn:nbn:de:bvb:473-irb-112551x
DOI: https://doi.org/10.20378/irb-112551

Publication List

The publications listed below are the result of the research carried out in this thesis titled "Advances in Bayesian Demographic Forecasting"

1. Goes, J. (2024). Bayesian forecasting of mortality rates for small areas using spatiotemporal models. *Demography*, *61*(2), 439–462. https://doi.org/10.1215/00703370-11212716 accepted version and published.

2. Goes, J., Barigou, K., & Leucht, A. (2025). Bayesian mortality modelling with pandemics: A vanishing jump approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *74*(4), 1150–1182. https://doi.org/10.1093/jrsssc/qlaf018 accepted version and published.

3. Goes, J., & Engelhardt, H. (2026). Probabilistic population forecast for small regions, *Demographic Research*, accepted version, not yet published.

Acknowledgments

First of all, I would like to express my sincerest gratitude to my supervisor, Prof. Dr. Anne Leucht, who at all times offered her support, guidance and help whenever I needed it. She always had an open door, took time for my many questions, and never once seemed bothered by them. Furthermore, I want to thank her for creating a wonderful environment at the Professorship of Mathematics for Economic Sciences, of which I was glad to be a part. She always prioritized the best interests of everyone, which allowed me to focus on obtaining my doctorate.

In addition, I want to thank Prof. Dr. Henriette Engelhardt-Wölfler for introducing me to the fascinating subject of statistical demography. Her expertise in the field and deep knowledge of the community were invaluable to my research. Furthermore, I would like to thank Prof. Dr. Timo Schmid for many insightful discussions, both professional and personal. Moreover, I would like to express my gratitude to Dr. Martin Messingschlager for encouraging me to work at the Institute of Statistics, which ultimately set me on this path. Lastly I want to thank Prof. Karim Barigou for giving me the wonderful opportunity to visit the University of Laval for a highly productive four-week research stay, which resulted in a joint paper and many great memories.

I am deeply grateful to all my colleagues, as well as the HiWi's at the Institute of Statistics, who made going to work something I genuinely looked forward to. In particular, the coffee breaks after lunch are memories I will fondly remember in the future. A special thanks goes to Guy Brunotte and Florian Scholze, who, despite my frequent interruptions, always took the time to offer their help. Additionally, I want to thank Florian Meinfelder for being not only a great teacher – who introduced me to the topic of Bayesian statistics – but also an even greater colleague, whom I could always ask for advice and occasionally discuss the performance of our favorite football team. Lastly, I want to thank my office mate Michael Mühlbauer, without whom the last five years would have been only half as fun, and who I could always count on to cheer me up when things weren't working and I couldn't figure out why. Finally, I want to thank Christine and Sabine whose organization and efficiency ensure that everything runs smoothly around here. Without their knowledge and hard work, nothing would get done.

Finally I want to thank all my friends for their support during these last years. I am forever grateful for my brothers, Max and Paul, for my Dad for always believing in me, and my Mum for her unconditional support.

Contents

Introduction

Demographic forecasting is a critical tool in today's society. It includes forecasts of population, fertility, mortality, migration, labor force participation, housing demand, and more. Users of these forecasts – typically governments or organizations – require them to be as precise as possible, including a realistic depiction of their underlying uncertainty. Accurate forecasts are essential for anticipating future needs and demands. For example, they are used to plan new schools, infrastructure projects, and nursing homes. The demographic forecasts in question may be local, or national and to provide the necessary detail, they are often disaggregated by age and sex. Perhaps the most important and often the most sought after demographic forecast concerns the prediction of future populations. Any population of interest – whether local or national – has three components of population change, namely mortality, fertility and net-migration. Thus, to predict future population values one can forecast each of these components separately and then appropriately combine them. This approach is called cohort-component method and was mathematically formalized by Leslie (1945). Other methods exist, for example, extrapolation methods of population totals, however the cohort-component method is now the most widely used for generating future population values (Preston et al., 2000).

The demographic literature distinguishes between population *projections* and population *forecasts*. Population projections using the cohort-component method typically define deterministic scenarios based on combinations of mortality, fertility, and migration. If each variable has, for example, three variants, this results in nine different combinations, from which medium, high, and low outcomes are then selected. As such, population projections are essentially calculations that illustrate the future development of a population under specific assumptions, and may be seen as 'what-if' scenarios. While often regarded as a best guess of the future, they do not allow for any probabilistic statements, making it impossible to tell which of these scenarios is more likely. Furthermore, such variants are internally inconsistent from a statistical point of view. For example, if the high-low variant of each input covers the 95%-probability interval, the resulting high-low variants for population size will not cover a 95%-probability (Lee, 1998). Unfortunately, many statistical agencies continue to produce deterministic population projections, such

as the one from the Federal Statistical Office of Germany (Statistisches Bundesamt), which includes a total of 21 variants (Statistisches Bundesamt, 2022). On the contrary, a population forecast is an estimate of the future population derived from a statistical model that uses historical data to predict future trends. A population forecast can be generated through a probabilistic implementation of the cohort-component method, which accounts for component uncertainty by assigning each input a statistical model with associated probability distributions, resulting in probabilistic forecasts of the future population. This produces internally consistent forecasts, allows for the generation of scenarios of varying probability – including a best guess usually in the form of a median forecast – and also informs about the degree of uncertainty. While the difference between a projection and forecast may sound clear in theory, users often have difficulty differentiating between the two, especially when population projections are published by statistical agencies. Put in another way: "a demographer makes a projection, and his reader uses it as a forecast" (Keyfitz, 1972, p. 353).

Various studies have shown that the accuracy of population forecasts decreases with the forecast horizon and is typically higher for larger populations than for smaller ones. Additionally, these forecasts are usually accompanied by a significant degree of uncertainty (Keilman, 2019). Therefore, users of population forecasts should be informed about the expected accuracy (or inaccuracy). Otherwise, they may develop a false sense of security, assuming that the projected population is certain to occur. Lee (1998, p. 156) even states that, "it is generally agreed that demographers have an responsibility to indicate how certain or uncertain their forecasts may be." This agreement has led to a change of culture in the demographic community. In the past, population projections were the standard tool of statistical agencies for predicting future population values. However, population forecasts have become increasingly popular. For instance, the United Nations (UN) now produces population forecasts every two years for all countries of the world, and in 2015, they adopted a fully probabilistic approach (see United Nations (2024) for an overview on their methodology). Additionally, extensive research has focused on improving population forecasts using the cohort-component method, by developing new methods for probabilistic predictions of its inputs: mortality, fertility, and net-migration (cf. Raftery & Ševčíková, 2023). This thesis adds to the ongoing research with the goal of developing new Bayesian methods for forecasting mortality, fertility, and net-migration that can then be used to obtain accurate forecasts of age-specific population on the national and regional scale.

The thesis is structured in three chapters. The first chapter focuses on obtaining precise forecasts of regional mortality. Mortality forecasting has received a great deal of attention in both the demographic and actuarial science literature (see, for example, Booth and Tickle (2008) for a review), including the seminal work by Lee and Carter (1992), known

as the Lee-Carter (LC) model, and the Age-Period-Cohort (APC) model by Hobcraft et al. (1982), which remain the standard approaches for generating future mortality forecasts. However, most research in the past has focused on predicting mortality rates on a national level. Forecasting mortality rates at a regional level has only recently received more attention in the literature. Borrowing ideas from spatial statistics, Chapter 1 extends the LC and APC models by a regional component that captures spatial correlation, thereby improving predictive accuracy on a regional level. To make the predictions more robust, a Bayesian version of stacking is considered using leave-future-out validation.

In 2020, the world was unexpectedly impacted by the COVID-19 pandemic, which rapidly spread across countries and led to a sharp increase in mortality rates, resulting in a significant number of excess deaths in Europe (Pizzato et al., 2024). Chapter 2 focuses on developing new statistical models that are able to account for large, unexpected jumps in mortality rates as those caused by pandemics. Existing models only allow for transitory jumps that affect a single period. However, there is no literature on estimating mortality models with jumps that have an effect over a small number of periods, as typically observed in pandemics. By introducing the concept of correlated mortality jumps – characterized by a high initial impact that gradually vanishes – the LC model is extended to allow for a more reasonable description of the mortality rates during pandemics resulting in a superior model fit.

The thesis concludes with Chapter 3, which focuses on predicting future age-specific populations for small subnational areas, using regions in Upper Franconia, Bavaria as an example. Age-specific population forecasts are derived by first predicting age-specific mortality, fertility, and net-migration. These are then combined using a probabilistic implementation of the cohort-component method. Regional mortality forecasts are obtained using the methods described in Chapter 1. For regional fertility, the LC model is extended with an age-region interaction, while net-migration counts are forecasted using skewed error terms. Additionally, migration patterns are estimated using a Dirichlet regression.

Chapter 1

Bayesian Forecasting of Mortality Rates for Small Areas Using Spatiotemporal Models [1]

Abstract

Estimation and prediction of subnational mortality rates for small areas are essential planning tools for studying health inequalities. Standard methods do not perform well when data are noisy, a typical behavior of subnational datasets. Thus, reliable estimates are difficult to obtain. I present a Bayesian hierarchical model framework for prediction of mortality rates at a small or subnational level. By combining ideas from demography and epidemiology, the classical mortality modeling framework is extended to include an additional spatial component capturing regional heterogeneity. Information is pooled across neighboring regions and smoothed over time and age. To make predictions more robust and address the issue of model selection, a Bayesian version of stacking is considered using leave-future-out validation. I apply this method to forecast mortality rates for 96 regions in Bavaria, Germany, disaggregated by age and sex. Uncertainty surrounding the forecasts is provided in terms of prediction intervals. Using posterior predictive checks, I show that the models capture the essential features and are suitable to forecast the data at hand. On held-out data, my predictions outperform those of standard models lacking a regional component.

Keywords: Mortality forecasting, Subnational estimation, Spatiotemporal models, Stacking, Bayesian hierarchical models

1.1 Introduction

Rapid aging of the population in Western countries has brought additional challenges for government and health agencies and led to an increased demand for accurate and reliable estimates of age-specific mortality rates, including life expectancy. Precise predictions of

[1] Accepted version of article published in: Goes, J. (2024). Bayesian forecasting of mortality rates for small areas using spatiotemporal models. *Demography*, *61*(2), 439–462. https://doi.org/10.1215/00703370-11212716

future life expectancy are particularly crucial for healthcare, pension plans, retirement funds, aged care, and the life insurance industries. Moreover, to study and uncover health disparities within a county, reliable estimates of mortality rates at the subnational or even municipality level are essential. Such estimates enable researchers to reveal heterogeneity within a population and allow policymakers to incorporate sensible regional policies.

Mortality estimation has become more sophisticated, yet subnational estimation still remains a challenge for multiple reasons: When data are disaggregated by age, sex, region, and time, the population count is usually small and it is common to observe zero cell counts for certain age groups. For example, in many regions of Bavaria, Germany, it is typical to observe zero deaths among children aged 5 to 10 in a given year. In such cases, traditional life table approaches break down, yielding mortality rate estimates of zero and, in turn, infinite life expectancy. Moreover, stochastic variation in subnational data is higher, making observations more noisy. Oftentimes, cells are combined—that is, they are aggregated into superregions until counts are large enough and random variation becomes less predominant (Ezzati et al., 2008; Murray et al., 2006). Other problems consist of shorter time series and erratic trends (Wilson et al., 2022).

In recent decades, direct estimates using life tables have been replaced with new methods producing probabilistic forecasts. Some of the more popular approaches include the well-known Lee–Carter model (LC) by Lee and Carter (1992); its cohort extension, the Renshaw–Habermann model (RH) by Renshaw and Haberman (2006); and the Age–Period–Cohort (APC) model by Hobcraft et al. (1982). All of these models were developed using a frequentist approach, but in recent years, Bayesian adaptations have also been proposed. The interested reader is referred to Bijak and Bryant (2016) for an overview and history of Bayesian methods in demography. In this article, I focus on forecasting mortality rates for all age groups at a subnational level using Bayesian implementations of the popular APC and RH models. Furthermore, these models are extended with spatially structured effects, making them more suitable for subnational mortality forecasting and improving prediction accuracy.

Significant improvements have been made in subnational mortality estimation, with Bayesian hierarchical models, in particular, showing promise. These models smooth estimates by pooling strength across dimensions such as age, time, and sex, making them well-suited for limited or sparse data. Applications in demography include the interesting approach of Alexander and Alkema (2022) and Alexander et al. (2017) using principal components and a Bayesian APC implementation by Bryant and Zhang (2016). While these approaches incorporate a region-specific effect, they do not account for spatial correlation, where neighboring regions are more similar than distant regions. Other successful, albeit different, methods for estimation of subnational mortality rates include, for

example, an application of the TOPLAS relational model (Rau & Schmertmann, 2020; Schmertmann & Gonzaga, 2018) and the application of Taylor's lay by Yang et al. (2022).

The use of spatial statistics allows for explicit incorporation of spatial correlation to the regional component. Xu et al. (2014) employed a Bayesian Poisson linear mixed model with and without a spatially structured effect to estimate regional child mortality, revealing that the spatial model outperformed the nonspatial counterpart in terms of in-sample fit. Consequently, a growing literature has examined the use of Bayesian spatial models for the estimation of subnational mortality rates (Congdon, 2014; Mercer et al., 2015; Ocaña-Riola & Mayoral-Cortés, 2010; Wakefield et al., 2019). However, all of the approaches focus on inference rather than forecasting.

The aims of this article are twofold. First, I propose to forecast mortality rates for small areas using Bayesian hierarchical models with the inclusion of a random effect capturing spatial heterogeneity. I demonstrate that this can be seamlessly integrated into the LC framework. The incorporation of spatial components into the LC family has not yet been broadly applied, to the best of my knowledge, although it has been introduced to the classical APC model. I do not consider other mortality models, such as the popular Cairns–Blake–Dowd model by Cairns et al. (2006). This model is appropriate for higher ages only and my goal is prediction for all age groups. For a detailed overview on methods for mortality modeling, see Booth and Tickle (2008) and references therein.

In addition, I focus on forecasting and its performance. The accuracy with and without the addition of a spatial component is compared by calculating multiple performance measures, including scores for the assessment of probabilistic forecasts. I then check how much gain in accuracy can be achieved. Hereby, I show that the inclusion of a correlated spatial effect increases prediction accuracy substantially and argue that it should become standard procedure if the goal is to forecast demographic rates for small areas. Second, I introduce existing methods from the Bayesian literature for the evaluation and assessment of the proposed models to the demographic literature. With the help of posterior predictive checks, I demonstrate that my models are adequate in describing the observed features of the data and, hence, are suitable for the task of prediction. Lastly, I follow the ideas of Barigou et al. (2023) and use stacking to aggregate forecasts by various models. Like any combination approach, stacking incorporates model uncertainty into the prediction problem. In a situation where it can be assumed that none of the models in question perfectly describes the true data-generating process (a realistic scenario), or when different models are best at describing separate parts of the data, stacking is appropriate and offers an intriguing, robust alternative that is more protected against model misspecification.

1.2 Data

For estimation of death rates, counts of deaths as well as population are needed. Data are available for regions in Bavaria, the second largest state of Germany in terms of population, and are provided by the Bayerisches Landesamt für Statistik (Bavarian Statistical Institute). The data are publicly available and can be downloaded from GENESIS, the database of the Bavarian Statistical Institute. The datasets consist of the total number of deaths (Bayerisches Landesamt für Statistik, 2022b) as well as population counts (Bayerisches Landesamt für Statistik, 2025) disaggregated by age, sex, region, and year. Age is given in groups of five years except for the first two and the last: age 0, ages 1-4, then five year age groups from 5-10 up to 90-95 and 95+, resulting in a total of $X = 21$ age groups. The data is given for $T = 17$ years, from 2001 to 2017, for a total of $R = 96$ regional districts in Bavaria. Hence, there are a total of $N = X \cdot T \cdot R = 21 \cdot 17 \cdot 96 = 34\,272$ death rates to be estimated per sex. Out of the 34 272 cells for each sex, there are 7 187 (11.1 %) and 4 806 (7.45 %) zero death counts for females and males, respectively. Summed up over age, the 96 regions range in population count from around 18 000 to 750 000 (for one sex). The lowest cell count in terms of population is four for males and 33 for females.

1.3 Methods

I use a hierarchical Bayesian modeling approach, where the counts are assumed to be Poisson distributed. This is in line with most Bayesian implementations of APC or LC models (Bryant & Zhang, 2016; Pedroza, 2006; Wiśniowski et al., 2015). Alternatively, one may also assume that deaths are sampled from a binomial distribution (e.g., Congdon, 2014). Even though data are available for both sexes, I do not attempt to model them together, meaning there will be a separate model for males and females, which is not unusual in the demographic literature (e.g., Alexander et al., 2017).

Let $y_{x,t,r}$ denote the death counts of age group $x = 1, \ldots, X$ at time $t = 1, \ldots, T$ in region $r = 1, \ldots, R$ with $\boldsymbol{y} = (y_{1,1,1}, \ldots, y_{X,T,R})^\top$. Moreover, assume that $y_{x,t,r}|M_{x,t,r} \sim \text{Poi}(E_{x,t,r}M_{x,t,r})$, where $M_{x,t,r}$ denotes the underlying mortality rate scaled to $E_{x,t,r}$, that is the person-years of exposure or the population exposed to that risk. Since the observation on population is given at the end of the period, person-years lived is approximated by

$$E_{x,t,r} = \frac{G_{x,t,r} - G_{x,t-1,r}}{\log \frac{G_{x,t,r}}{G_{x,t-1,r}}},$$

where $G_{x,t,r}$ denotes the population at age x, time t and region r (Preston et al., 2000, p. 15). Given the Poisson family, a log link connects the mortality rate to the linear predictor with $\eta_{x,t,r} = \log(M_{x,t,r})$.

I extend the classical APC model by inclusion of a spatial term. The linear predictor is as follows

$$\eta_{x,t,r} = \mu + \alpha_x + \kappa_t + \gamma_k + \phi_r + \varepsilon_{x,t,r}. \tag{1.1}$$

Here, the parameter ϕ_r denotes the regional effect capturing spatial heterogeneity and μ the global intercept, that is, the average log-mortality rate. Parameters α_x, κ_t and γ_k denote the respective age, time and cohort effects. The age effect is a static function describing age related effects, while κ_t denotes the evolution of mortality over time. The cohort effect γ_k represents the life long effects specific to a birth cohort and its index k is a function of age and period. If the intervals are of different lengths, that is, if the age intervals are M times wider than the period intervals, the cohort index is given by $k = M(X - x) + t$, with $k \in \{1, \ldots, M(X-1) + T\}$ (Heuer, 1997). In my case, $M = 5$. Lastly, the error term $\varepsilon_{x,t,r}$ accounts for overdispersion.

In the RH model, the linear predictor is extended with the addition of the same spatial component,

$$\eta_{x,t,r} = \alpha_x + \beta_x^{(1)}\kappa_t + \beta_x^{(2)}\gamma_k + \phi_r + \varepsilon_{x,t,r}. \tag{1.2}$$

Here, the parameter α_x denotes the average log-mortality rate at age x. This static age function models the general shape of mortality by age. The parameter κ_t estimates the global change over time and γ_k the global effect of cohort k. The parameters $\beta_x^{(1)}$ and $\beta_x^{(2)}$ measure the response to changes of κ_t and γ_k, respectively, at age x. The regional effect ϕ_r captures spatial dependencies. An error term accounts for overdispersion.

To achieve identifiability, some restrictions have to be imposed on the parameters. For all models, the regional parameter needs to be constrained depending on the type of model implemented. Details are given later. For the LC family, I invoke the typical constraints, that is, $\sum_x \beta_x^{(1)} = \sum_x \beta_x^{(2)} = 1$ and $\sum_t \kappa_t = \sum_k \gamma_k = 0$ (Renshaw & Haberman, 2006).

For the APC subfamily, the matter is more complex, and the problem of identification has been thoroughly discussed in the literature (see Smith and Wakefield (2016) for a review). To start, some constraints have to be imposed on the main effects for the intercept μ to be identifiable. This can be achieved by either a corner constraint, that is, to fix a parameter value to zero (e.g. $\kappa_1 = 0$), or a sum-to-zero constraint. I follow the latter, more popular choice in the literature, and set $\sum_t \kappa_t = \sum_k \gamma_k = \sum_x \alpha_x = 0$ (e.g., Riebler & Held, 2017). The sum-to-zero constraints give identifiability of the intercept but do not solve the identification problem per se. Because of the linear dependence between the age, period, and cohort effect, the overall mean is invariant to the addition of a constant, meaning that the full set of effects is not identifiable (Smith & Wakefield, 2016). Unfortunately, there is no solution to this problem. For identification of single

effects, further constraints are needed, such as assuming that two period effects are equal. Alternatively, one of the age, period, or cohort effects can be removed from the model (Smith & Wakefield, 2016). Since the main focus is on forecasts and not estimation of parameters, these problems can be neglected and additional constraints avoided as long as future mortality rates are identified. Kuang et al. (2008) demonstrated that this depends on how future period and cohort effects are extrapolated. Examples include a random walk with drift and a random walk of order two. Alternatively, a zero mean forecast may be used (Smith & Wakefield, 2016).

1.3.1 Prior Specifications

For the prior specifications, so-called weakly informative priors are used instead of vague or uninformative priors. That is, the prior should rule out unreasonable values but not be too restrictive that it precludes values that might make sense. Since the priors on the parameters are set on the log-scale and the mortality rate $M_{x,t,r} \in (0, 1]$, realistic values for the linear predictor $\eta_{x,t,r}$ range from around –15 to 0[2]. Hence, it does not make sense to set an uninformative prior for the intercept, such as $\mu \sim \mathcal{N}(0, 10000)$, since most of the probability mass of that prior results in impossible values for mortality rates. Therefore, I chose more restrictive priors, $\mu \sim \mathcal{N}(-5, 5)$. The overdispersion effect is modeled using a centered independent normal prior $\varepsilon_{x,t,r} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$.

The age effect typically shows dependencies between adjacent age groups and can consequently be modeled using time-series methods. This is a typical approach in the demographic and APC literature (Alho & Spencer, 2005; Congdon, 2014; Riebler & Held, 2017). Here, the age effects are assumed to follow a random walk model of order 2 with Gaussian error, that is,

$$\begin{aligned} \Delta^2 \alpha_x &= \nu_x \\ (\alpha_x - \alpha_{x-1}) - (\alpha_{x-1} - \alpha_{x-2}) &= \nu_x, \end{aligned}$$

with $\nu_x \overset{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$. This can be stated as

$$\alpha_x | \alpha_{x-1}, \alpha_{x-2}; \sigma_a^2 \sim \mathcal{N}\left(2\alpha_{x-1} - \alpha_{x-2}, \sigma_\alpha^2\right).$$

The first two age effects are modeled separately using centered normal priors with the same variance σ_α^2. For the additional age effects $\beta_x^{(i)}$ of the RH model, I choose a Dirichlet prior, because of the implied sum-to-one constraint, as done by Barigou et al. (2023), with $\beta_x^{(i)} \sim \text{Dirichlet}(1, \ldots, 1)$, for $i = 1, 2$.

[2] A value of the linear predictor of -15 results in a mortality rate of $\exp(-15) = 0.000\,000\,03$ which is less than the lowest predicted age-specific mortality rate of the official UN forecast for the year 2100 (see https://population.un.org/wpp/ for data and results)

The time effect is modeled as a random walk with drift, as proposed by Lee and Carter (1992), with

$$\kappa_t = c + \kappa_{t-1} + \omega_t, \tag{1.3}$$

where $\omega_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\kappa^2)$ and $c \sim \mathcal{N}(0, 2)$.

The cohort effect is modeled via an unstructured normal prior, where $\gamma_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_\gamma^2)$. Lastly, all standard deviations are given half t-distributed hyperpriors with five degrees of freedom.

Spatially structured priors

Given a set of observations at R different spatial units (i.e., regions), spatial interaction between pairs of units may be modeled via spatially structured priors. These priors smooth estimates by pooling information from neighboring regions. Hence, strength is borrowed locally instead of globally. There are several options for the selection of spatially dependent priors. Here, I use spatial autoregressive models, more precisely the BYM2 model as proposed by Riebler et al. (2016), an extension of the famous Besag–York–Mollie model (BYM) of Besag et al. (1991).

Let $\boldsymbol{u} = (u_1, \ldots, u_R)^\top$ denote a spatially structured effect that follows an intrinsic conditional autoregressive model (ICAR) belonging to the class of conditional autoregressive models (CAR). Spatial interaction between pairs of units is given by a conditional distribution $p(u_r|u_{-r})$, that is, the distribution of a regional effect u_r given all other regions u_{-r}. The ICAR prior for u_r is given by

$$p(u_r|u_{-r}; \sigma_u^2) \sim \mathcal{N}\left(\frac{\sum_{r\neq j} w_{rj} u_j}{\sum_{r\neq j} w_{rj}}, \frac{\sigma_u^2}{\sum_{r\neq j} w_{rj}}\right), \tag{1.4}$$

where σ_u^2 is the variance parameter and w_{rj} denote symmetric weights, indicating spatial dependence between two regions r and j with $r, j \in \{1, \ldots, R\}$. The weights w_{rj} are assumed to be binary indicators of adjacency. Thus, $w_{rj} = 1$ if two regions r and j are neighbors, that is they share a common border, and $w_{rj} = 0$ otherwise.

In the above parameterization of the ICAR model, the effectu_r is normally distributed with mean equal to the average of its neighbors. It should be noted, that the distribution in Eq. (1.4) is improper as it defines only the differences between pairs of units and not its overall level. This distribution cannot be used as a model for the data but it can still serve as a prior (Banerjee et al., 2015, p. 155). The spatially structured effect is therefore constrained to sum to 0, that is, $\sum_r u_r = 0$.

The ICAR prior of Eq. (1.4) does not allow for spatially unstructured variation. To address this issue, the BYM2 model combines both a spatially structured and an un-

structured component. In addition, a single variance parameter σ_ϕ^2 is placed on the combined components, while a mixing parameter $\rho \in [0,1]$ accounts for the amount spatial or nonspatial variation. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_R)^\top$, then the BYM2 prior is given by

$$\boldsymbol{\phi} = \sigma_\phi \left(\sqrt{1-\rho}\, \boldsymbol{v}^* + \sqrt{\rho}\, \boldsymbol{u}^* \right).$$

Here, $\boldsymbol{v}^* = (v_1^*, ..., v_R^*)^\top$ denotes a scaled spatially unstructured effect, with, $v_r^* \overset{iid}{\sim} \mathcal{N}(0,1)$ for all $r = 1, \ldots, R$, and $\boldsymbol{u}^*$ a scaled ICAR model with geometric mean of its marginal variances approximately equaling 1 (Riebler et al., 2016). If $\rho = 0$, the model reduces to a spatially unstructured effect, while $\rho = 1$ leads to a fully spatially structured effect. The mixing parameter ρ is given a Beta$(0.5, 0.5)$ hyperprior.

1.3.2 Forecasting

In a Bayesian setting, forecasting is part of the estimation process and carried out via evaluation of the posterior predictive distribution. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \mu, c)^\top$ denote the parameters of interest, then the h-step ahead predictive distribution is

$$p(y_{x,T+h,r}|\boldsymbol{y}) = \int p\left(y_{x,T+h,r}|\boldsymbol{\theta}, \boldsymbol{y}\right) p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}.$$

In our setting, the integral is analytically not tractable but can be approximated using Monte Carlo simulations. Having obtained S posterior draws of all model parameters $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)})$ the posterior predictive distribution $p(y_{x,T+h,r}|\boldsymbol{y})$ can be approximated by

$$p\left(y_{x,T+h,r}|\boldsymbol{y}\right) \approx \frac{1}{S} \sum_{s=1}^{S} p\left(y_{x,T+h,r}|\boldsymbol{y}, \boldsymbol{\theta}^{(s)}\right).$$

For prediction of future time points, new parameter values need to be generated of effects that vary over time, that is, the time, cohort effect, and future error terms. The other parameters are time invariant, so they can be drawn from the joint posterior. Suppose you are given S posterior draws for all model parameters, then the sth $(s = 1, \ldots, S)$ value of the forecasted h-step ahead mortality rate can be calculated as follows:

Step 1: For $h = 1, \ldots, H$, draw $c^{(s)}$ from the posterior and generate new $w_{T+h}^{(s)}$ by sampling from $\mathcal{N}(0, \sigma_\kappa^2)$. Use both to generate new values of $\kappa_{T+h}^{(s)}$ repeatedly using Eq. (1.3).

Step 2: For $h = 1, \ldots, H$, generate $\gamma_{M(X-x)+(T+h)}^{(s)}$ by drawing from $\mathcal{N}(0, \sigma_\gamma^2)$.

Step 3: For $h = 1, \ldots, H$ generate new $\varepsilon_{x,T+h,r}^{(s)}$ by drawing from $\mathcal{N}(0, \sigma_\varepsilon^2)$.

Step 4: Get the sth posterior draw of all remaining parameters and plug all values into Eq. (1.2), respectively, Eq. (1.1) to compute $\log\left(\widehat{M}^{(s)}_{x,T+h,r}\right)$.

Step 5: Exponentiate $\log\left(\widehat{M}^{(s)}_{x,T+h,r}\right)$ to obtain future mortality rates.

After having obtained forecasts of mortality rates $\widehat{M}_{x,T+h,r}$, they can be converted into deaths $\hat{y}_{x,T+h,r}$. For each posterior predictive draw, sample from a Poisson distribution to obtain $\hat{y}^{(s)}_{x,T+h,r}$, with $\hat{y}^{(s)}_{x,T+h,r} \sim \text{Poi}\left(E_{x,T+h,r}\widehat{M}^{(s)}_{x,T+h,r}\right)$. Mean forecasts of other quantities of interest are approximated via Monte Carlo simulations. It should be noted that future values of deaths can only be predicted for known $E_{x,T+h,r}$. Hence, for future time periods without known exposure, that is, 2018 onward, only rates are predicted, which in turn may be used as inputs for life table methods, such as future life expectancy at birth.

1.4 Parameter Estimation

In Bayesian statistics, inference on the parameters is achieved via evaluation of the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. Analytical solutions for high-dimensional distributions are oftentimes not available. Hence Markov chain Monte Carlo (MCMC) methods are employed. Here, the probabilistic modeling software `Stan` (Stan Development Team, 2024c), a tool for Hamiltonian Monte Carlo (HMC), was used to produce samples from the posterior distribution. Stan can be accessed through the interface `rstan` in `R` (Stan Development Team, 2024a). HMC, a variant of MCMC, uses Hamiltonian dynamics for the proposition of a new state. It avoids random walk behavior and exhibits less autocorrelation, making it more efficient by requiring fewer iterations than standard MCMC methods. An introduction can be found in Neal (2011). The code and data for all models are available at https://github.com/goesj/BavarianMortality. Information regarding the size of the chains and run time may be found in the appendix in section A.1

Convergence of parameters was checked using the build-in diagnostic measures of `rstan`, which are described in Vehtari et al. (2021). Those quantities are split-$\hat{R}$ as well as tail and bulk effective sample size (ESS). The measures are calculated for each scalar quantity of interest, that is, all parameters as well as hyperparameters.

The first measure, split-$\hat{R}$, is a variant of the estimated scale reduction factor $\hat{R}$. This factor compares the total variance relative to the within-chain variance, indicating how much the posterior samples vary within each chain. The idea is as follows: if the chains have not mixed well, the variance of all simulations together should be higher than the individual chains. Values of $\hat{R}$ near 1 suggest that the total variance is similar to the within-chain variance, indicating that all of the M chains are exploring the same distribution. Vehtari et al. (2021) identified cases in which the traditional $\hat{R}$ fails to detect nonconvergence and,

therefore, proposed a split-$\hat{R}$ variant. In this variant, rank-normalized values are used, and each chain is split in half, resulting in twice as many chains. For computation details, see Vehtari et al. (2021). The authors argue for a very tight threshold and propose to use the sample only if split-$\hat{R} < 1.01$.

The second measure, ESS, incorporates information about the degree of autocorrelation of the posterior draws, as samples from HMC are not independent. Each draw is dependent on the value of the last iteration. High autocorrelation leads to slow mixing and possibly nonconvergence. The amount of autocorrelation is used to calculate the effective sample sizeESS, a scale-free measure for diagnosing the sampling efficiency. Bulk-ESS refers to the effective sample size using rank-normalized draws, while tail-ESS denotes the minimum of the effective sample sizes of the 5% and 95% quantiles. Computational details can be found in Vehtari et al. (2021), where the authors propose to use the sample only if both ESS measures are greater than 100 times the number of chains.

1.5 Stacking

Instead of using a single model for prediction, the idea of model combination can be applied. Here, point predictions of multiple models are averaged according to a specific weight. The simplest technique assigns equal weights to all models, while stacking (Wolpert, 1992) weighs each model according to their predictive performance. An implementation by Yao et al. (2018) generalized the idea of stacking to combine Bayesian predictive distributions rather than just mean forecasts. Suppose there are K models $\mathcal{M} = (M_1, \ldots, M_k)$ each with a predictive distribution $p(.|M_k)$. The object of stacking K predictive distributions from models $\mathcal{M}$ is to find the distribution in the convex hull $\mathcal{C} = \left\{ \sum_{k=1}^{K} a_k \cdot p(.|M_k) : \sum_k a_k = 1, a_k \geq 0 \right\}$ that is best according to some criterion. Here, I follow the approach to that outlined by Yao et al. (2018) and use the negative log score, that is, the negative logarithm of the predictive density, to define the optimality criterion. With respect to the chosen score, stacking finds the predictive distribution that is closest to the true data-generating process (Yao et al., 2018). Moreover, it tackles the model selection problem in a data driven way. Similar to Barigou et al. (2023), the weights are derived with the leave-future-out (LFO) predictive density, that is, the predictive density of future observations. In the approach by Yao et al. (2018), stacking weights are derived using the leave-one-out predictive density, although owing to the dependence structure in my data, the likelihood cannot be factorized, making cross-validation procedures more complex.

Let a_k denote the weights associated with each mortality model $M_k \in \mathcal{M}$. The weights

can be obtained by solving the following optimization problem

$$\min_{a \in \mathcal{A}^K} \frac{1}{N} \sum_{x,T+h,r} \mathcal{D}\Bigg[\sum_{k=1}^{K} a_k \Bigg(\frac{1}{S} \sum_{s=1}^{S} p(y_{x,T+h,r} | \boldsymbol{\theta}_k^{(s)}, \boldsymbol{y}, M_k) \Bigg) \Bigg],$$

where $\mathcal{D}$ denotes a suitable scoring function, for example, the negative log score [3] and

$$\mathcal{A}^K = \Bigg\{ (a_1, \ldots, a_k)^\top \in [0,1]^K : \sum_{k=1}^{K} a_k = 1 \Bigg\}.$$

The stacked h-step ahead predictive distribution is then

$$\begin{aligned} \hat{p}(y_{x,T+h,r} | \boldsymbol{y}) &= \sum_{k=1}^{K} a_k \, p(y_{x,T+h,r} | \boldsymbol{y}, M_k) \\ &= \sum_{k=1}^{K} \Bigg(\frac{1}{S} \sum_{s=1}^{S} a_k \, p(y_{x,T+h,r} | \boldsymbol{\theta}_k^{(s)}, \boldsymbol{y}, M_k) \Bigg). \end{aligned}$$

1.6 Model Checks and Evaluation

In the following we use the term *estimated* when referring to in-sample, that is historic rates and *predicted* or *forecasted* when referring to out-of-sample rates.

1.6.1 Model Checks

First, to assess whether the additional spatially structured parameter is necessary for describing the data at hand, I follow the ideas outlined by Banerjee et al. (2015, pp. 75–76) and compute one of the standard measures of spatial association, Geary's C (Geary, 1954), for each age and time period. As recommended, Geary's C is not used as a test of spatial significance but rather as an exploratory tool. However, the distributional assumptions underlying Geary's C assume constant mean and variance across regions. Both assumptions are not fulfilled in our setting, as the mean and the variance depend on the population size. Waller and Gotway (2004, p. 235) proposed to replace the death counts $y_{x,t,r}$ by a standardized value $z_{x,t,r}$ under the constant risk hypothesis

$$z_{x,t,r} = \frac{y_{x,t,r} - M_{x,t} E_{x,t,r}}{\sqrt{M_{x,t} E_{x,t,r}}},$$

[3] Note, that I have defined the log score to be negative logarithm of the predictive density. Hence, it becomes a minimization instead of a maximization problem (Yao et al., 2018).

where $M_{x,t} = \frac{\sum_{r=1}^{R} y_{x,t,r}}{\sum_{r=1}^{R} E_{x,t,r}}$. Using $z_{x,t,r}$, I can compute Geary's C as

$$C_{x,t} = \frac{(R-1)\sum_r \sum_j w_{rj}\,(z_{x,t,r} - z_{x,t,j})^2}{2\big(\sum_{r\neq j} w_{rj}\big)\sum_r (z_{x,t,r} - \bar{z}_{x,t})^2},$$

where w_{rj} denote binary indicators of association as described above. The measure has a mean of 1 in the null model, while values close to 0 or 2 suggest positive or negative spatial association, respectively (Waller & Gotway, 2004). Hence, values deviating from 1 indicate spatial association in some form. Using the data at hand, I found an average value over ages for C of around 0.85 for most years, indicating moderate positive spatial association and suggesting the need for a spatially structured effect. A box plot of all values can be found in the appendix in section A.7.

Bayesian model validation employs the use of posterior predictive checks. The idea is intuitive: if the model is a good fit, then I should be able to use it to generate new data that closely resemble the observed data. I simulate multiple replicate datasets by drawing samples from the posterior predictive distribution and compare those to the observed data. Systematic differences suggest that the model does not capture the essential features of the data and should be improved upon. Let $\boldsymbol{y_{rep}}$ denote the replicated data. It can be created using the posterior predictive distribution

$$p(\boldsymbol{y_{rep}}|\boldsymbol{y}) = \int p(\boldsymbol{y_{rep}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}.$$

Additionally, a test statistic $T[\boldsymbol{y}]$ or $T[\boldsymbol{y}, \boldsymbol{\theta}]$ can be defined that is a scalar summary of the data. It may be used to compare the replicated with the observed data. There is no clear guidance regarding which test statistic to calculate. In the present case, I compute the proportion of predicted zeros for each age group. This test statistic can be used to check for zero inflation of the data (e.g., Neelon et al., 2013). Let $\boldsymbol{y^{(x)}}$ denote the vector of deaths for each age group over region and time, then

$$T[\boldsymbol{y^{(x)}}] = \sum_{t,r} \mathbb{1}(y_{x,t,r} = 0)/(T \cdot R),$$

where $\mathbb{1}$ denotes the indicator function. The test quantity can be evaluated for each replicated data set to obtain a vector of $T[\boldsymbol{y_{rep}^{(x)}}, \boldsymbol{\theta}]$ and construct a histogram thereof. Ideally, the test statistic of the original data should lie somewhere in the middle of the histogram. Discrepancies, that is, if the observed data test statistic lies at the outer ends of the histogram, suggest poor fit because the replicate distribution cannot reproduce the observed data.

1.6.2 Model Evaluation

For model evaluation, I follow the standard procedure in time-series modeling. I estimate and forecast the model on a subset of the available data, called the training set, and assess forecast accuracy on the held-out or test data. Unfortunately, data are not abundant, so only short-term forecasts may be evaluated. Let the training set consist of the first J years with $\boldsymbol{y_{1:J}} = (y_{1,1,1}, \ldots, y_{X,J,R})^\top$. The remaining $L = T - J$ years are considered as test set, with $l = J + 1, \ldots, J + L$.

After generating forecasts for all models, I assess their quality by comparing the predicted to the observed deaths using multiple performance measures. Point forecasts, specifically the mean forecasts of our predictive distribution, are evaluated using the mean absolute error (MAE) and the root mean squared error (RMSE). Let $y_{x,l,r}$ denote the actual observed value of the test set and $\widehat{\overline{M}}_{x,l,r}$ the mean of the posterior predictive distribution for the year $J + l$.

For sake of simplicity let $\widetilde{N} = X \cdot L \cdot R$. The measures are as follows

$$\text{MAE} = \frac{1}{\widetilde{N}} \sum_{x,l,r} \left| y_{x,l,r} - \widehat{\overline{M}}_{x,l,r} E_{x,l,r} \right|$$

$$\text{RMSE} = \sqrt{\frac{1}{\widetilde{N}} \sum_{x,l,r} \left(y_{x,l,r} - \widehat{\overline{M}}_{x,l,r} E_{x,l,r} \right)^2}.$$

I am interested not only in point forecasts, but also in the uncertainty surrounding them. For that, I need to evaluate the entire predictive distribution, which can be achieved using scoring rules. I follow the approaches outlined in Czado et al. (2009) by calculating the negative log score (LogS), the Dawid-Sebastiani score (DSS) proposed by Gneiting and Raftery (2007) as well as the ranked probability score (RPS). Similar to point measures, lower scores suggest a better fit. It should be noted that there are more scoring rules available. The interested reader is referred to Czado et al. (2009). The use of scoring rules for comparison of probabilistic forecasts within a demographic application is not new (e.g., Barigou et al., 2023; Keilman, 2020), though not common despite its relevance.

The negative log score is given by the negative logarithm of the predictive density (lpd) evaluated at the observation, defined as

$$-\text{lpd} = -\log p(y_{x,l,r} | \boldsymbol{y_{1:J}}) = -\log \int p(y_{x,l,r} | \boldsymbol{\theta}, \boldsymbol{y_{1:J}}) p(\boldsymbol{\theta} | \boldsymbol{y_{1:J}}) d\boldsymbol{\theta}.$$

In practice, it is calculated using draws from the posterior:

$$\text{LogS}_{x,l,r} = -\widehat{\text{lpd}} = -\log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_{x,l,r}|\boldsymbol{\theta}^{(s)})\right).$$

Second, the DSS of an observation is defined as follows:

$$\text{DSS}_{x,l,r} = \left(\frac{y_{x,l,r} - \mu_{x,l,r}}{\sigma_{x,l,r}}\right)^2 + 2\log(\sigma_{x,l,r}),$$

where $\mu_{x,l,r} = \mathbb{E}(y_{x,l,r})$ denotes the mean and $\sigma_{x,l,r} = \sqrt{\text{Var}(y_{x,l,r})}$ the standard deviation of the predictive distribution. Using the law of iterated expectations and the law of total variance, it follows that $\mu_{x,l,r} = E_{x,l,r} \cdot \mathbb{E}(M_{x,l,r})$ and $\sigma^2_{x,l,r} = E_{x,l,r} \cdot \mathbb{E}(M_{x,l,r}) + E^2_{x,l,r} \cdot \text{Var}(M_{x,l,r})$. The proof can be found in the appendix in section A.2. Lastly, the ranked probability score is given by

$$RPS_{x,l,r} = \sum_{z=0}^{\infty} \left[P_z - \mathbb{1}(y_{x,l,r} \leq z)\right]^2,$$

where $(P_z)_{z=0}^{\infty}$ denotes the distribution function of $p(y_{x,l,r}|\boldsymbol{y_{1:J}})$. Evaluation of the distribution function in my setting is analytically not tractable. On the other hand, S draws from the posterior predictive are available, which can be used to compute the empirical distribution function. Then, the RPS reduces to

$$\text{RPS}_{x,l,r} = \frac{1}{S}\sum_{s=1}^{S} \left|\hat{y}^{(s)}_{x,l,r} - y_{x,l,r}\right| - \frac{1}{2S^2}\sum_{s=1}^{S}\sum_{j=1}^{S} \left|\hat{y}^{(s)}_{x,l,r} - \hat{y}^{(j)}_{x,l,r}\right|$$

(Jordan et al., 2019).

In assessing the predictive distribution, I compare the mean scores of all models. This is particularly relevant for Poisson-distributed variables concerning overdispersion. Two models with different predictive distributions could yield the same mean forecast, but only one correctly estimates its surrounding uncertainty by indicating a lower score.

Another problem concerning discrete data is the width of the (sample) quantiles and thus prediction intervals (PI). For a forecasted random variable $\hat{Y}_{T+h}$, a PI denoted as $[y_l, y_u]$ with coverage level p_{Cov} is given by

$$P(y_l \leq \hat{Y}_{T+h} \leq y_u) \geq p_{Cov}. \tag{1.5}$$

Standard Bayesian practice computes PIs by calculating sample quantiles of the posterior predictive distribution and then uses these as the respective values for y_l and y_u. The PIs

are usually chosen to be equal-tailed, that is, $\frac{(1-p_{Cov})}{2}\%$ of the distribution's probability lies on either side of the bounds y_l and y_u. Hereafter I will call this procedure classical PI. When the sample size is large, the sample quantiles denote a good approximation and the estimated PI interval has a coverage close to or exactly equal to the nominal value p_{Cov}.

For discrete-valued data (e.g., Poisson) calculation of Eq. (1.5) comes with its share of difficulties. The reasons are twofold. First, one can select values for y_l and y_u to exactly attain the desired coverage level p_{Cov} for continuous random variables. However, this is often not possible for discrete data because the distribution function has jumps. As a result, the PIs usually have a larger coverage than intended. This effect can become even more severe if one naively relies on symmetric intervals. Second, empirical quantiles might be inconsistent estimators for the true quantiles, see e.g. Jentsch and Leucht (2016) for an example.

While the inconsistency of empirical quantiles cannot be resolved, I can still aim to find PIs with the desired coverage level p_{Cov}. Homburg et al. (2021) introduced an algorithm that finds a set of discrete PIs $[y_l^*, y_u^*]$, denoted as coherent PIs, based on sample quantiles that do not have to be equal-tailed or central like classical PIs, while attempting to minimize exceeding the coverage level in Eq. (1.5). The coherent PI is more flexible and similar in style to a highest posterior density region, which denotes an interval that contains $p_{Cov}\%$ of the posterior probability. Both the coherent PI and classical PI are equal for symmetric distributions, such as a normal distribution. While the Poisson distribution is almost symmetric with high means, for smaller means, the converse holds, and the distribution becomes right-skewed. Given the data, it makes sense to use a more flexible approach since the mean values range from low to high, and the coherent PI can provide both a central PI and a highest density PI. Moreover, the coherent PI has performed better in terms of "average exceedance", that is, the average amount by which the PI exceeded the true coverage level, than equal-tailed PIs based on empirical quantiles in a small simulation study I created. Therefore, PIs for forecasted deaths are calculated using the algorithm of Homburg et al. (2021). Details about the algorithm and simulation study are given in the appendix in section A.3.

1.7 Application to Real Mortality Data

I applied all of the described models to the Bavarian dataset split by sex. I estimated the proposed models with and without the addition of a regional component. All models are denoted with their usual abbreviations, including the type of regional component after an underscore. Hence, APC_BYM2 stands for an APC model with the addition of the BYM2 model as a spatial component.

1.7.1 Model Checks

Given the above priors, the measures split-$\hat{R}$ and tail as well as bulk ESS suggested convergence for all models and parameters. Moreover, for both the RH_BYM2 and APC_BYM2, the posterior predictive checks did not reveal any major discrepancies. First, no consistent deviations can be found when plotting the replicate against the observed data distribution. Some fairly extreme observations are identified where the observed data lie at the outer ends of the replicated data distribution, though all observed values can be recreated using the replicate draws. Exemplary results for both females and males can be found in the appendix in section A.7. Second, the zero inflation check suggests no need for a different model. For both males and females, the models predict a few to many zeros for the middle age groups, though the derivations are not extreme. This aspect can be observed for both the APC and RH models using different types of priors. This seems to be a special feature of the data and should be investigated separately, though it does not indicate the need for a zero-inflated model. The details may be found in section A.7 in the appendix for females and males. I therefore conclude that both the RH_BYM2 and APC_BYM2 are suitable for explaining the data.

1.7.2 Model Evaluation

For the evaluation of the predictive performance, I split the data into test and training sets. The test set comprises all observations from the years 2001 to 2014, while the training set includes death counts for all age groups and regions from 2015 onward. After estimation, mortality rates for 2015 to 2017 were forecasted and transformed into deaths using draws from the Poisson distribution. The predicted deaths for all regions and age groups were compared with the observed ones of the test data. Predictive accuracy was evaluated using the forecast measures and scoring rules as described above. For the calculation of stacking weights, the training data are divided into two parts. The years 2001 to 2010 are used for training, while 2011 to 2014 are used for validation, that is, the derivation of weights. The split between training (fitting) and test (prediction) is summarized in Table 1.1.

Table 1.1: Fitting, validation and prediction periods of all approaches in the model evaluation scheme

Model	Fitting	Validation	Prediction
Age-Period-Cohort	2001 - 2014		2015-2017
Renshaw-Haberman	2001 - 2014		2015-2017
Stacking	2001 - 2010	2011-2014	2015-2017

Female Data

Looking at the models by themselves, the APC_BYM2 is best in point prediction accuracy as well as scoring rules. The RH_BYM2's performance is slightly worse, albeit by a small margin. When comparing the results of both models with their counterparts lacking the regional component, it becomes evident that the addition of a regional component enhances predictive power across the board, lowering both the scores and point measures. In terms of point forecasts, a reduction of approximately 10% in MAE and 20% in RMSE was observed when comparing the APC_BYM2 model to the standard APC model, lacking the regional component. Similar findings were observed for the RH_BYM2 model, for which the addition of the regional component reduces the MAE by approximately 7% and the RMSE by slightly more than 20% when compared with the RH model Results for the female data is given in Table 1.2.

Looking at the models by themselves, the APC_BYM2 is best in point prediction accuracy as well as scoring rules. The RH_BYM2's performance is slightly worse, albeit by a small margin. When comparing the results of both models with their counterparts lacking the regional component, it becomes evident that the addition of a regional component enhances predictive power across the board, lowering both the scores and point measures. In terms of point forecasts, a reduction of approximately 10% in MAE and 20% in RMSE was observed when comparing the APC_BYM2 model to the standard APC model, lacking the regional component. Similar findings were observed for the RH_BYM2 model, for which the addition of the regional component reduces the MAE by approximately 7% and the RMSE by slightly more than 20% when compared with the RH model. Results for the female data are given Table 1.2.

Table 1.2: Evaluation of out-of-sample forecast for 2015-2017 of female models

Model	Mean LogS	Mean DSS	Mean RPS	MAE	RMSE	Coverage
APC	2.32	2.88	3.15	4.48	12.11	0.85
APC_BYM2	2.29	2.82	2.90	4.14	**9.45**	0.85
RH	2.32	2.88	3.22	4.53	14.21	0.84
RH_BYM2	2.29	2.83	2.99	4.21	11.20	0.84
Stacking	**2.28**	**2.81**	**2.88**	**4.12**	9.51	0.84

Notes: Values in bold denote the best of each column. LogS = log score. DSS = Dawid–Sebastiani score. RPS = ranked probability score. MAE = mean absolute error. RMSE = root mean squared error. APC = Age–Period–Cohort. BYM2 = Besag–York–Mollie 2. RH = Renshaw–Haberman.

Looking at the results split by years, I observe inconsistency in that no single model is superior over each year. Forecasts of the RH_BYM2 model for the year 2015 perform better than those of the APC_BYM2 model with regard to scoring rules and point measures. However, for the years 2016 and 2017, the APC_BYM2 clearly outperforms the RH_BYM2. Results are given in section A.6 in the appendix. For this reason, I conclude

that there is no single best model available for different parts of the data or forecast horizons. Depending on the training set or time of forecast, one model may be better than another.

With stacking one would hope to find a robust predictor that performs well regardless of time and forecast horizon. Using the years 2011 to 2014 for the derivation of weights, I observe the approach heavily favoring the APC_BYM2 model, resulting in the following weights: .871 (APC_BYM2) and .129 (RH_BYM2). Using the derived weights, stacked predictions were calculated for 2015 to 2017. The results are shown in Table 1.2. Looking at the stacked forecasts, I notice very comparable, though slightly better, values with regard to the scoring rules and MAE. However, there is a slight decline in RMSE compared with that of the APC_BYM2 model. When split by years, the robustness of the stacking approach becomes noticeable, consistently either the best or second-best performing approach. Details are given in A.6 in the appendix.

Coverage was estimated at the 80% level. A coverage greater than the nominal value for all models was obtained, yet still close to the 80% level. This is potentially due to both the discreteness of the random variable and the great amount of zeros within the test set, inflating the measure. By excluding the zero observations, coverage dips very close to 80% for all models, which is a satisfactory mark.

Male Data

For the male data, I observe similar findings, in that the addition of a regional component aids in forecasting. Moreover, the APC_BYM2 model is superior to all other models with regard to point measures and scoring rules. Complete results are given in Table 1.3. Interestingly, the stacking approach does not heavily favor the APC_BYM2 model even though the model's performance in 2015 to 2017 is by far the best. Using the years 2011 to 2014 as validation, that is, derivation of weights, the RH_BYM2 model received the highest weight at .546, compared to .454 for the APC_BYM2. This result aligns with the observed behavior of the female data, for which different models excel at forecasting separate parts of the data. Thus, despite the RH_BYM2's struggles in the test period, it performs best in the validation period. The stacked predictions rank second best in all categories except for RMSE (see Table 1.3), an unsurprising result considering the performance of the RH_BYM2 model in the test period. Results split by years may be found in A.6 in the appendix. Coverage at the 80% level is a little higher than the nominal value of 80% for all models, similar to that for the female data.

The results for males are interesting in the sense that the data at hand seem to exhibit some special features that no single model can predict best. Different models favor different time periods, which are unknown beforehand to the user. Thus, it might be that

Table 1.3: Evaluation of out-of-sample forecast for 2015-2017 of male models. Value in bold denotes the best of the column.

Model	Mean LogS	Mean DSS	Mean RPS	MAE	RMSE	Coverage
APC	2.59	3.45	3.48	4.95	11.69	0.84
APC_BYM2	**2.53**	**3.34**	**3.00**	**4.20**	8.43	0.84
RH	2.69	3.89	3.57	5.00	11.52	0.80
RH_BYM2	2.66	3.89	3.14	4.33	8.48	0.80
Stacking	2.60	3.58	3.05	4.27	**8.36**	0.80

Notes: Values in bold denote the best of each column. LogS = log score. DSS = Dawid–Sebastiani score. RPS = ranked probability score. MAE = mean absolute error. RMSE = root mean squared error. APC = Age–Period–Cohort. BYM2 = Besag–York–Mollie 2. RH = Renshaw–Haberman.

the chosen test period supports a single model that may not be suitable for forecasting another period adequately.

1.7.3 Results and Forecasts

For the actual forecasts of mortality rates, that is, predictions for 2017 onward, I chose to implement the stacking approach. Particularly for the male dataset, I cannot be certain that a single model performs best over all years. There is substantial variation in performance across different time periods. Hence, selecting a single model could easily result in misspecification, as one does not know how the data will behave in the future. Stacking, on the other hand, incorporates model uncertainty into the prediction problem to obtain more robust forecasts. It was observed that these predictions are either the best or, at the very least, close to the best, making stacking suitable for the task at hand. Additionally, even though the models in question describe the data fairly well, it is reasonable to believe that the complex dynamics of the true data-generating process are not exactly equal to any one of the models within our pool, rendering no single model perfectly suitable. Weights for the final, stacked predictions were estimated by training the data on the years 2001 to 2010 and forecasting on 2011 to 2017. Hereafter, they are referred to as final stacking weights to avoid confusion. The procedure is outlined Table 1.4.

Table 1.4: Fitting, validation and prediction periods of all approaches

Model	Fitting	Weight Derivation	Prediction
APC	2001 - 2017		2018 - 2030
RH	2001 - 2017		2018 - 2030
Stacking	2001 - 2010	2011 - 2017	2018 - 2030

Notes: APC = Age–Period–Cohort. RH = Renshaw–Haberman.

The final weights for the female data are .906 for the APC_BYM2 and .094 for the RH_BYM2 model. These weights are similar to those received in the model evaluation

period, with the weight of the RH_BYM2 decreasing a bit. For males, the final weights are .630 (APC_BYM2) and .370 (RH_BYM2). In comparison with the model evaluation period, the weight distribution has changed, resulting in the APC_BYM2 model being favored.

Following the estimation and prediction of age-specific death rates, life table methods were employed to transform rates into life expectancy at birth, hereafter referred to as LE. For the last observed year – 2017 – the map in Fig. 1.1 shows the mean LE estimates of all regions split by sex. Panel a displays those of females and panel b displays those of males. For both sexes, the same pattern emerges—a north–south and east–west pattern is clearly visible. This pattern aligns with that of other social indicators such as GDP and available wealth, as shown in Bayerisches Landesamt für Statistik (2022a, p. 13). In addition to point estimates, PIs can be easily obtained by transforming each set of age-specific mortality rates drawn from the posterior predictive distribution into LE. PIs are then calculated based on the LE draws. In 2017, the region with the highest rank for both sexes is Starnberg, with a mean LE of 84.98 years (80% PI: [84.80, 85.19]) for the total population and 81.91 years (80% PI: [81.73, 82.09]) for females and males, respectively. For the year 2017, the LE estimates for all regions, including PIs, are plotted in the appendix in section A.7. The plots are sorted by mean LE.

When moving from estimation to forecasting, there is a sharp increase in uncertainty, a typical pattern in time-series analysis. In comparing the LE estimate for 2017 (in-sample fit) with the prediction for 2018 (out-of-sample fit), I observe a widening of the surrounding intervals. In general, mean predictions of LE are expected to increase linearly over time (owing to the linear prior) for both sexes. However, the forecasted growth of mean LE is lower than the estimated in-sample growth. In other words, LE is expected to increase less in the future than it has in the past. Since there are no spatiotemporal interaction terms within the model, the time effect is global, and its behavior is constant for all regions. Exemplary for Bavaria as a whole, LE forecasts for the region of Bamberg, City can be found in Figure Fig. 1.2. The complete results for all other regions is available in a Shiny app online. [4]

Additionally, when examining Fig. 1.2, it becomes apparent that the LE of males has increased more than that of females, especially in the early 2000s. Plotting the difference in estimated mean LE in 2001 versus the mean predicted LE in 2030, I observe the same effect for all regions – a decrease in the gender gap. Results are shown in Fig. 1.3. This is in line with the current findings of other countries (e.g., Sundberg et al., 2018). Interestingly, the LE increase from 2001 to 2030 is higher in regions where LE had been lower in 2001 and vice versa, especially for males.

[4] https://github.com/goesj/BavarianMortality

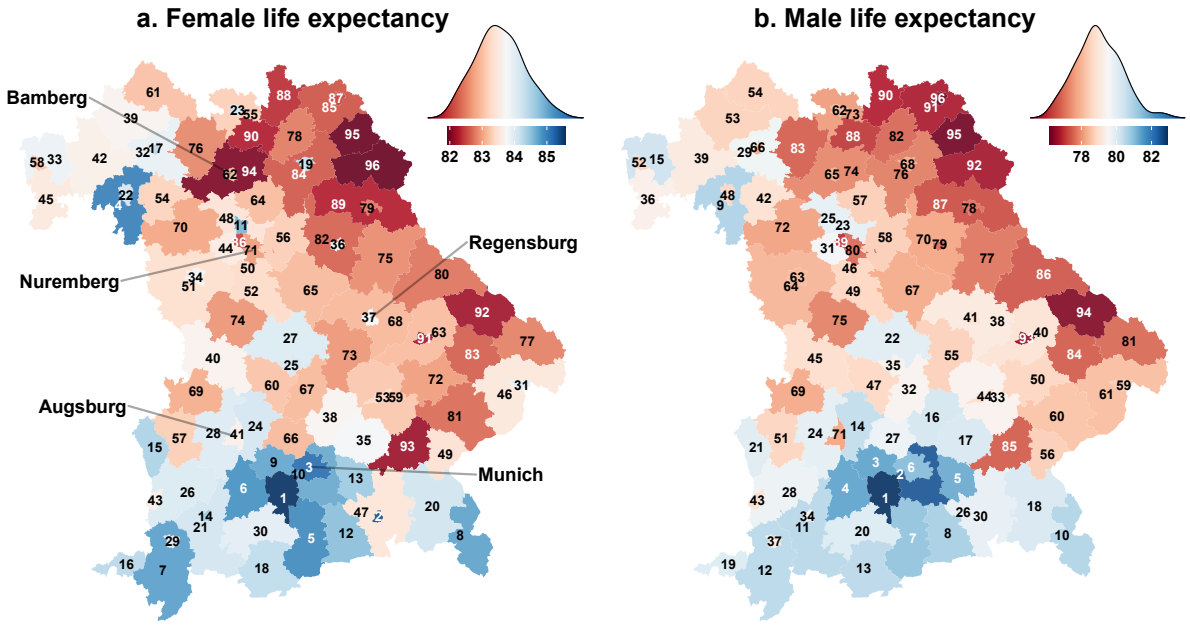


Figure 1.1: Mean estimates of life expectancy for all regions of Bavaria in 2017, using the stacking approach. Female estimates are plotted in panel a, and values for males are shown in panel b. Red areas denote regions with lower life expectancy, while blue areas denote regions with higher life expectancy. Within each region, the subsequent rank is plotted: 1 indicates the region with the highest life expectancy, and 96 indicates that with the lowest life expectancy. The density scales represent the kernel density estimate of all mean predictions.

Lastly, Fig. 1.2 shows a relatively sharp increase in LE in 2014 with a dip the following year. This phenomenon can be observed in most European countries and has been thoroughly discussed in the demographic literature (e.g., Luy et al., 2020). I can therefore conclude that this is due to some special feature of the data and does not indicate problems with the model.

1.8 Extensions

All presented models can be extended to include interaction terms between various components. While an age–time interaction for the APC model seems like an obvious choice, other interactions are also possible, such as a region–time interaction effect that may help to reveal different mortality trends in separate areas. Additionally, instead of estimating males and females separately, I could model both sexes jointly. For example, the approach by Wiśniowski et al. (2015) incorporated multivariate priors, such as vector autoregressive models for the time effect, allowing for correlations between sexes. Alternatively, as proposed by Bryant and Zhang (2016) and Zhang and Bryant (2020), one could add a sex-specific intercept with joint time and age effects.

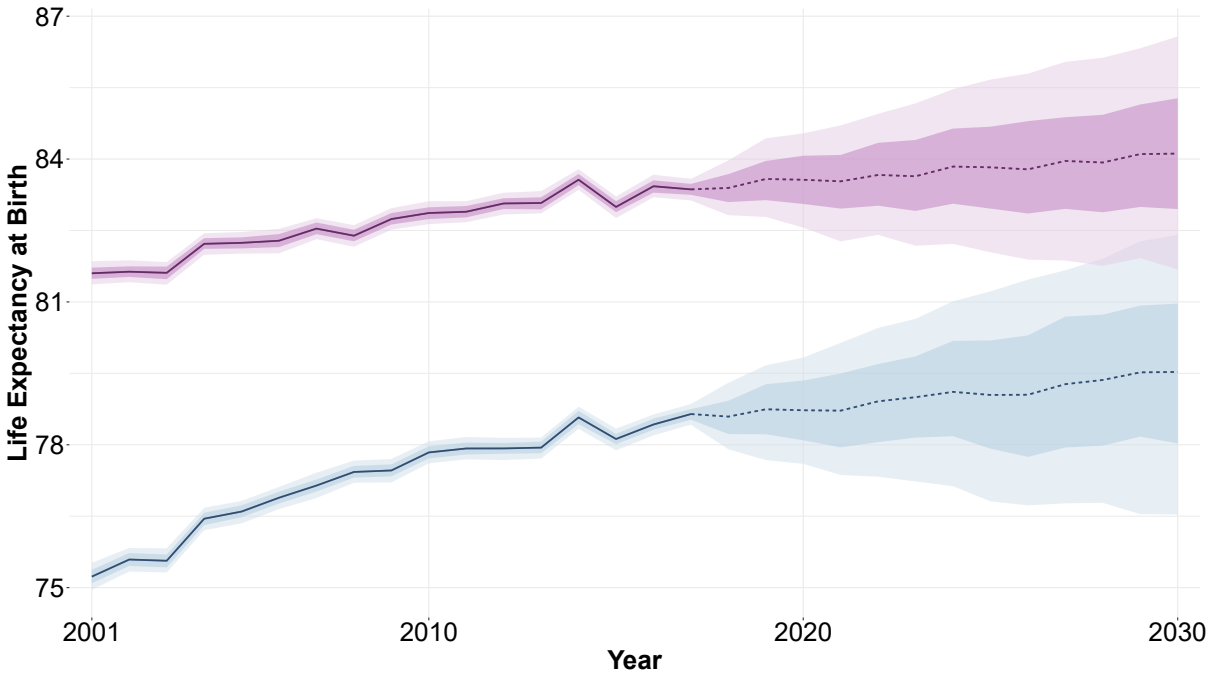


Figure 1.2: Prediction and estimation of life expectancy for the region of Bamberg, City, using the stacking approach. The solid lines denote the mean estimation for 2001 to 2017, and the dashed lines denote the mean prediction for 2018 to 2030. The shades of purple and blue represent PI for females and males respectively. Darker shades show 50% PIs, while lighter shades show 80% PIs.

Moreover, the population or exposure at risk in my analysis is seen as deterministic. This may be considered an unrealistic assumption, especially for small population sizes. A possible solution treats the exposure not as fixed but as a parameter itself, though this is not further explored. On the other hand, adding more uncertainty into the estimation process potentially increases the prediction intervals even further. In addition, underlying causes of deaths were not analyzed. The models presented can easily be adjusted to estimate cause-specific death rates as done, for example, in Richardson et al. (2013). Alternatively, an additional cause of death specific effect may be added.

Typically, for the assessment and comparison of predictive performance, leave-one-out-cross-validation (LOO-CV) is implemented. However, this approach is computationally expensive, especially in a Bayesian context, and oftentimes not feasible. Vehtari et al. (2017) developed an approach to approximate the leave-one-out (LOO) predictive density, but it relies on the factorization of the likelihood. For nonfactorizable models, such as time-series or spatial models, the situation is more complex. Given the time and regional dependence structures in the dataset, standard LOO-CV cannot be carried out as this dependence has to be accounted for. To properly account for the time-series structure, Bürkner et al. (2020) implemented the leave-future-out-cross-validation (LFO-CV) approach in a Bayesian context. LFO-CV for both model assessment and the calculation

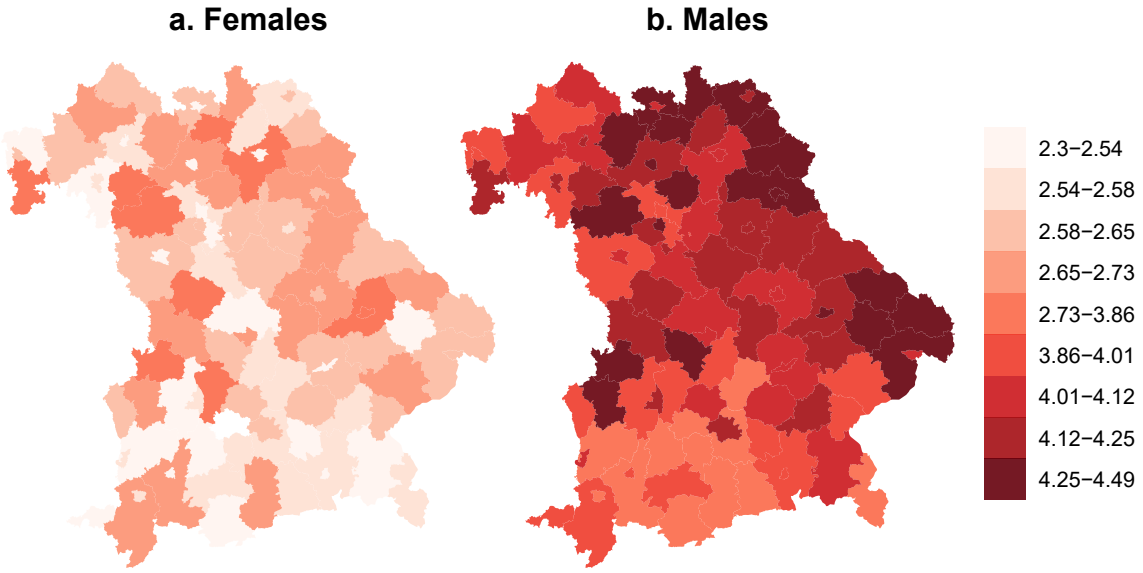


Figure 1.3: Differences in mean life expectancy between 2001 and 2030 for all regions in Bavaria, by sex, using the stacking approach. Panel a shows the difference for females and panel b shows that for males. Darker shades indicate regions with a greater increase in life expectancy.

of stacking weights, while also accounting for the regional structure, is an interesting direction for future research but is not further explored here.

The inclusion of covariates is another interesting line of direction. Multiple authors have found a negative relationship between income and mortality (e.g. Felice et al., 2016; Lorentzen et al., 2008). Integrating covariates into the model framework is straightforward and has the potential to significantly improve the predictive performance for each area. However, most covariates, such as GDP, are time dependent, meaning that they have to be forecasted as well. This is often challenging and can introduce potential biases, so I refrain from doing so. Moreover, all the unobserved spatial heterogeneity, such as the differences in regional income and infrastructure, is captured by the regional component, at least in theory.

An advantage of using Bayesian methods is the integration of informative priors arising from expert knowledge. Especially in sparse data settings, where the prior is given more weight, this may improve predictive accuracy. An interesting application in the field of demography is given in Billari et al. (2014). The incorporation of more informative priors, especially for the spatial effect, constitutes a possible extension to this framework.

Lastly, I have not looked at or incorporated the COVID-19 pandemic into the forecasts. It is expected that the pandemic will influence death rates in the short term, especially for older ages. Here, the pandemic is ignored in that I assume the effect of COVID-19

does not affect the long-term forecasts. I assume that the death rates for the years 2020 to 2023 experience a one-time effect and then tend back to their original, prepandemic path, a potentially problematic assumption. One possible solution by Liu and Li (2015) introduced single period jumps into the LC model. These jumps may be incorporated into my model framework and is potential for further research.

1.9 Discussion

In this article I have presented an approach for forecasting age-specific mortality rates, including life expectancy, by region and sex. I have added a spatially structured effect to both the APC and the RH model that captures regional correlations. My Bayesian framework pools information across dimensions, allowing for estimation in a sparse data context. Measures of uncertainty are provided in terms of prediction intervals. With the automatic modeling software Stan, implementation is rather straightforward and does not involve deviation of complex full conditionals. To protect against model misspecification, I have implemented the stacking approach, where predictions of multiple models are weighted according to their past performance. This method offers robust predictions and even improved predictive accuracy for the female data. When tested against simpler models without a spatially structured effect, the method outperformed them when applied to real data for 96 regions in Bavaria, Germany. Especially for small areas, where random variation is high, pooling strength across geographic regions stabilizes predictions. In addition, I introduced a variety of techniques for model evaluation that are not commonly found in the demographic literature. Posterior predictive checks offer a means of detecting conflicts between the model and data, revealing potential needs for extending or modifying an existing model. I advocate for their use as a standard tool.

When considering the forecasted death rates, these findings reveal slightly lower values in terms of scoring rules and MAE for females compared with those for males (see Tables 1.2 and 1.3, respectively). Additionally, the models for males exhibit longer convergence times, and the parameter estimates for males indicate higher levels of uncertainty. Consequently, the resulting prediction intervals, particularly for out-of-sample forecasts, are wider in most regions. This holds true for both the APC_BYM2 and the RH_BYM2 model, as well as the stacking approach. The results suggest that either the models are better suited for forecasting female death rates – that is, they describe the underlying data-generating process more accurately – or the higher random variation in the male data makes it more challenging to obtain accurate predictions. In either case, obtaining precise estimates and forecasts of death rates for small areas, for both men and women, remains a challenging task.

Lastly, I have primarily focused on the mean of the posterior predictive distribution, but

care must be taken with this approach, as the sole focus on mean estimates has its faults. The wide prediction intervals, especially in 2030, indicate a great deal of uncertainty, meaning that the point forecasts should not be considered as precise values on which to draw fixed conclusions. In general, as forecasts extend further into the future, greater uncertainty leads to wider prediction intervals, a characteristic pattern of the random walk with drift model. Only for stationary autoregressive moving-average models do prediction intervals converge to a constant width. Most models with an underlining trend, to which random walks with drift belong to, have ever-increasing intervals. Thus, if one expects life expectancy to increase in the future, one must accept the increased uncertainty in terms of wider intervals. Policymakers, particularly for small areas, should not rely heavily on point estimates alone. The inherent uncertainty, which grows with time, underscores the need for caution even in interpreting in-sample estimates.

Acknowledgment

I gratefully acknowledges financial support by the Oberfrankenstiftung (grant FP01054). Moreover, I thank Karim Barigou, Henriette Engelhardt-Wölfler and Anne Leucht for valuable discussions on the topic. Finally, I would like to thank two anonymous reviewers and editors for helpful comments on an earlier version of this manuscript.

A Appendix

A.1 HMC Information

For each model, four parallel chains were constructed. This is the minimum amount of chains recommended by Vehtari et al. (2021), as multiple chains increase the likelihood of revealing multimodality or poor mixing (Vehtari et al., 2021). Running more chains does not necessarily increase computational burden, as chains can be let run parallel. The first 2000 respectively 2500 iterations were considered warm-up and discarded while the rest was used for inference. The warm-up phase in HMC is not equivalent to the burn-in period in MCMC. During the warm-up phase `Stan` tunes the algorithm, whereas in standard MCMC the burn-in is used to ensure that the sampler has reached the desired target distribution.

In Table A.1 information regarding the HMC Model parameters can be found.

Table A.1: HMC Sampling Information of all models

Sex	Model	Warm-Up	Iterations	Thin	Adapt Delta
Female	APC	2 000	4 000	4	0.81
Female	APC_BYM2	2 000	4 000	4	0.81
Female	RH	2 000	4 000	4	0.81
Female	RH_BYM2	2 000	4 000	4	0.81
Male	APC	2 500	5 000	5	0.81
Male	APC_BYM2	2 500	5 000	5	0.81
Male	RH	2 500	5 000	5	0.81
Male	RH_BYM2	2 500	5 000	5	0.81

A.2 Derivation of DSS Parameters

Let $\mathbb{E}(y_{x,t,r}) = \mu_{x,t,r}$, then the law of iterative expectations states

$$\mu_{x,t,r} = \mathbb{E}\big(\mathbb{E}(y_{x,t,r}|\eta_{x,t,r})\big) = \mathbb{E}(E_{x,t,r} \cdot M_{x,t,r}) = E_{x,t,r} \cdot \mathbb{E}(M_{x,t,r}).$$

Analogously, the variance $\sigma^2_{x,t,r} = \mathrm{Var}(y_{x,t,r})$ is given by the law of total variance as

$$\begin{aligned}\sigma^2_{x,t,r} &= \mathbb{E}\big(\mathrm{Var}(y_{x,t,r}|\eta_{x,t,r})\big) + \mathrm{Var}\big(\mathbb{E}(y_{x,t,r}|\eta_{x,t,r})\big) \\ &= \mathbb{E}(E_{x,t,r} \cdot M_{x,t,r}) + \mathrm{Var}(E_{x,t,r} \cdot M_{x,t,r}) \\ &= E_{x,t,r} \cdot \mathbb{E}(M_{x,t,r}) + E^2_{x,t,r} \cdot \mathrm{Var}(M_{x,t,r}).\end{aligned}$$

A.3 Coherent Prediction Interval

For sake of simplicity and readability, let $\hat{y}$ define a random variable with distribution equal to that of the posterior predictive distribution of a forecasted death count $p(y_{x,T+h,r})$. Moreover, let $0 \leq y_l \leq y_u$ denote the integer-valued bounds of respective prediction interval (PI) of $\hat{y}$. The calculation of coherent PI's of Homburg et al. (2021) for a target coverage level p_{Cov} is given as follows:

Step 1: First compute the largest integer $L \in \{0, 1, \dots\}$, such that $P(\hat{y} < L) \leq 1 - p_{Cov}$.

Step 2: Then, for all $l = 0, 1, \dots, L$, compute the smallest integer u, such that $P(l \leq \hat{y} \leq u) \geq p_{Cov}$.

Step 3: Among the $L + 1$ resulting PI's, choose the one(s) having minimal length.

Step 4: If there exist several PI's $[y_{l,i}, y_{u,i}]$ of minimal length, choose the one with the greatest coverage:

$$P(\hat{y} \in [y_l, y_u]) = \max_i P(\hat{y} \in [y_{l,i}, y_{u,i}])$$

Simulation Study

The performance of the coherent prediction interval (coherent-PI) was evaluated using a small simulation study. Here, we compared the the coherent-PI with the mid-quantile approach by Ma et al. (2011) (mid-PI), as well as the standard approach, that is taking empirical quantiles as the integer valued bounds of the PI, which we will denote as standard-PI. The details of the simulation study can be found in Algorithm 1. Note, that

Algorithm 1: Simulation Study

for $s = 1$ **to** S **do**
  for $m = 1$ **to** M **do**
    Draw $N = 1000$ times from $P_m \sim \text{Poi}(\lambda_m)$, with $\lambda_m = m$
    Using draws from P_m calculate the respective PI at a given coverage level p_{Cov}
    Create annother $N = 1000$ forecasts from $\hat{P}_m \sim \text{Poi}(\lambda_m)$
    Determine a coverage level c_m by checking how many of the N forecasted values lie within the estimated prediction interval
  end
  Compute a set of sample statistics from $\{c_1, \dots, c_M\}$
end

in our study we set $M = 500$ while repeating the entire process $S = 200$ times. After having obtained coverage levels we compute the same sample statistics as Homburg et al. (2021), that is

1. the "average exceedance": The average amount of exceeding the intended coverage level p_{Cov}, given by the mean of $c_m - p_{Cov}$ for all $c_m > p_{Cov}$.
2. the average coverage given level given by $\bar{c}$
3. the sample standard deviation of all c_m.

The results are given in form of box plots in Fig. A.1.

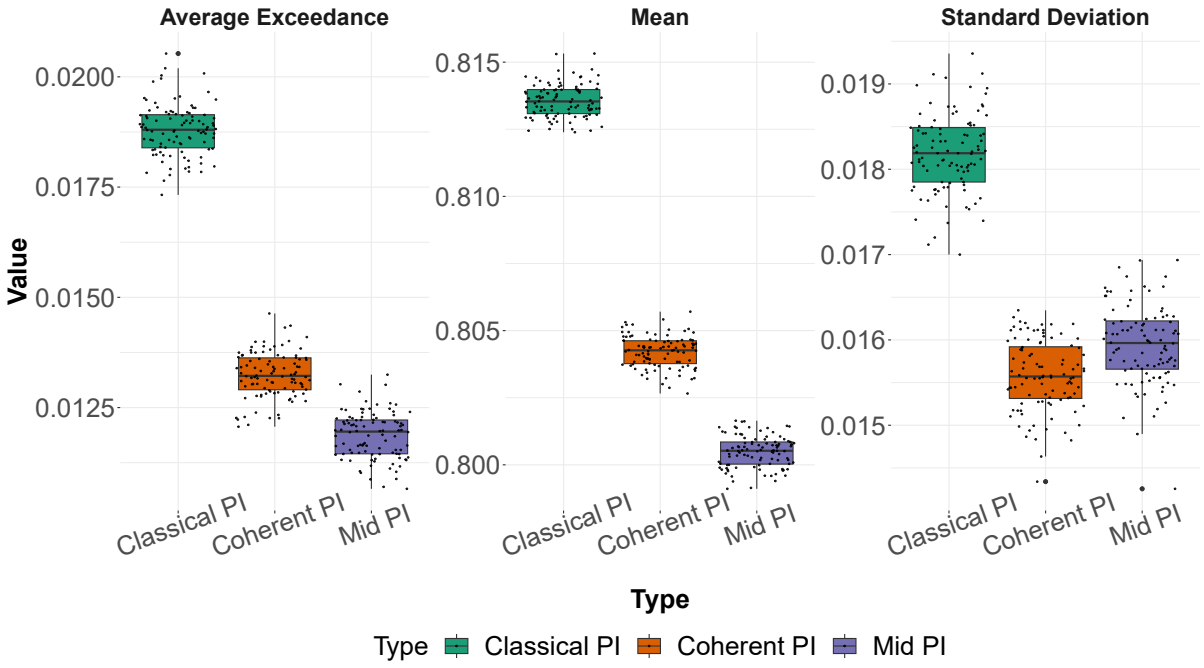


Figure A.1: Boxplot with results of the simulation study. The small black points denote the results of each sample statistic for each iteration by PI's type, while the boxplot is a summary of those points.

A.4 Additional Model Checks

To assess weather the spatial structured parameter is even necessary for describing the data at hand, we computed a measure of spatial association, Geary's C, for each age group and time. Results for both males and females can be found in Fig. A.2. Values diverting from one show spatial association of some form. Fig. A.2 shows a box plot of all ages groups for each year for both males and females. For every year, we observe a median value of Geray's C of below one, denoting positive spatial correlation.

A.5 Posterior Predictive Checks

To test the validity of our model we employed posterior predictive checks. Hereby, we generated some replicated data denoted $\boldsymbol{y_{rep}}$

$$p(\boldsymbol{y_{rep}}|\boldsymbol{y}) = \int p(\boldsymbol{y_{rep}}|\boldsymbol{\theta}) p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}.$$

The replicated data is then compared with the observed data to check for systematic differences. Exemplary results for females using the RH_BYM2 model and for males using the APC_BYM2 model are shown in Fig. A.3, respectively Fig. A.4.

Additionally, a test statistic $T[\boldsymbol{y}]$ or $T[\boldsymbol{y}, \boldsymbol{\theta}]$ can be defined that is a scalar summary of the data. Ideally, the test statistic of the original data should lie somewhere in the middle of the histogram. Discrepancies, that is if the observed data test statistic lies at the outer ends of the histogram, suggest poor fit because the replicate distribution cannot reproduce the observed data. Exemplary results for both females and males calculating the proportion of zeros using the APC_BYM2 model are given in Fig. A.5, respectively Fig. A.6.

PIT

In addition to posterior predictive checks, the probability integral transform (PIT) may be used as a diagnostic tool for calibration checks. A version for count data, called nonrandomized PIT, was proposed by Czado et al. (2009). If the observations of the held-out data are drawn from our predictive distribution, a desirable situation, the PIT has a uniform distribution. Hence, we may plot the PIT histogram and check for uniformity. U-shaped histograms indicate underdispersed predictions, while inversely U-shaped histograms suggest overdispersion.

After fitting the models on the test data, we can evaluate the PIT histogram of the predictive distribution. For all models, the PIT histogram showed good calibration. We also estimated the APC model of Eq. (1.1) without the overdispersion parameter $\varepsilon_{x,t,r}$ and plotted its PIT histogram. As expected, a U-shaped plot was observed suggesting underdispersion (see Fig. A.7).

A.6 Additional Tables

Table A.2: Evaluation of out-of-sample forecast for 2015-2017 of female models by Year. Value in bold denotes best of the column.

Jahr	Model	Mean Log	Mean DSS	Mean RPS	MAE	RMSE	Coverage
2015	APC	2.32	2.85	3.18	4.48	11.19	0.83
2015	APC_BYM2	2.29	2.80	2.97	4.26	8.74	0.84
2015	RH	2.30	2.82	3.07	4.25	11.80	0.84
2015	RH_BYM2	**2.28**	**2.78**	**2.90**	**4.11**	9.09	0.83
2015	Stacking	2.29	2.80	2.97	4.24	8.74	0.83
2016	APC	2.29	2.84	3.06	4.37	13.88	0.86
2016	APC_BYM2	2.26	2.78	2.75	3.97	**10.51**	0.87
2016	RH	2.30	2.87	3.24	4.54	16.27	0.84
2016	RH_BYM2	2.27	2.81	2.94	4.14	12.53	0.84
2016	Stacking	**2.25**	**2.76**	**2.75**	**3.96**	10.59	0.86
2017	APC	2.34	2.93	3.21	4.60	11.04	0.84
2017	APC_BYM2	2.31	2.87	2.96	4.19	**9.00**	0.85
2017	RH	2.34	2.95	3.34	4.79	14.20	0.84
2017	RH_BYM2	2.32	2.89	3.13	4.39	11.70	0.84
2017	Stacking	**2.31**	**2.86**	**2.94**	**4.16**	9.09	0.84

Table A.3: Evaluation of out-of-sample forecast for 2015-2017 of males models by Year. Value in bold denotes best of the column.

Jahr	Model	Mean Log	Mean DSS	Mean RPS	MAE	RMSE	Coverage
2015	APC	2.58	3.44	3.44	4.87	11.33	0.83
2015	APC_BYM2	**2.52**	**3.31**	**2.98**	**4.17**	**8.43**	0.83
2015	RH	2.64	3.66	3.52	4.94	11.16	0.80
2015	RH_BYM2	2.60	3.62	3.14	4.35	8.62	0.79
2015	Stacking	2.57	3.49	3.06	4.27	8.49	0.79
2016	APC	2.59	3.47	3.44	4.92	12.02	0.85
2016	APC_BYM2	**2.55**	**3.38**	**2.98**	**4.18**	8.60	0.85
2016	RH	2.80	4.46	3.59	4.97	11.60	0.81
2016	RH_BYM2	2.80	4.58	3.16	4.29	8.44	0.80
2016	Stacking	2.68	3.88	3.05	4.23	**8.37**	0.80
2017	APC	2.59	3.44	3.56	5.06	11.70	0.85
2017	APC_BYM2	**2.53**	**3.33**	**3.04**	**4.25**	8.26	0.87
2017	RH	2.62	3.56	3.61	5.10	11.78	0.81
2017	RH_BYM2	2.57	3.47	3.12	4.37	8.39	0.82
2017	Stacking	2.54	3.39	3.04	4.30	**8.23**	0.82

A.7 Additional Figures

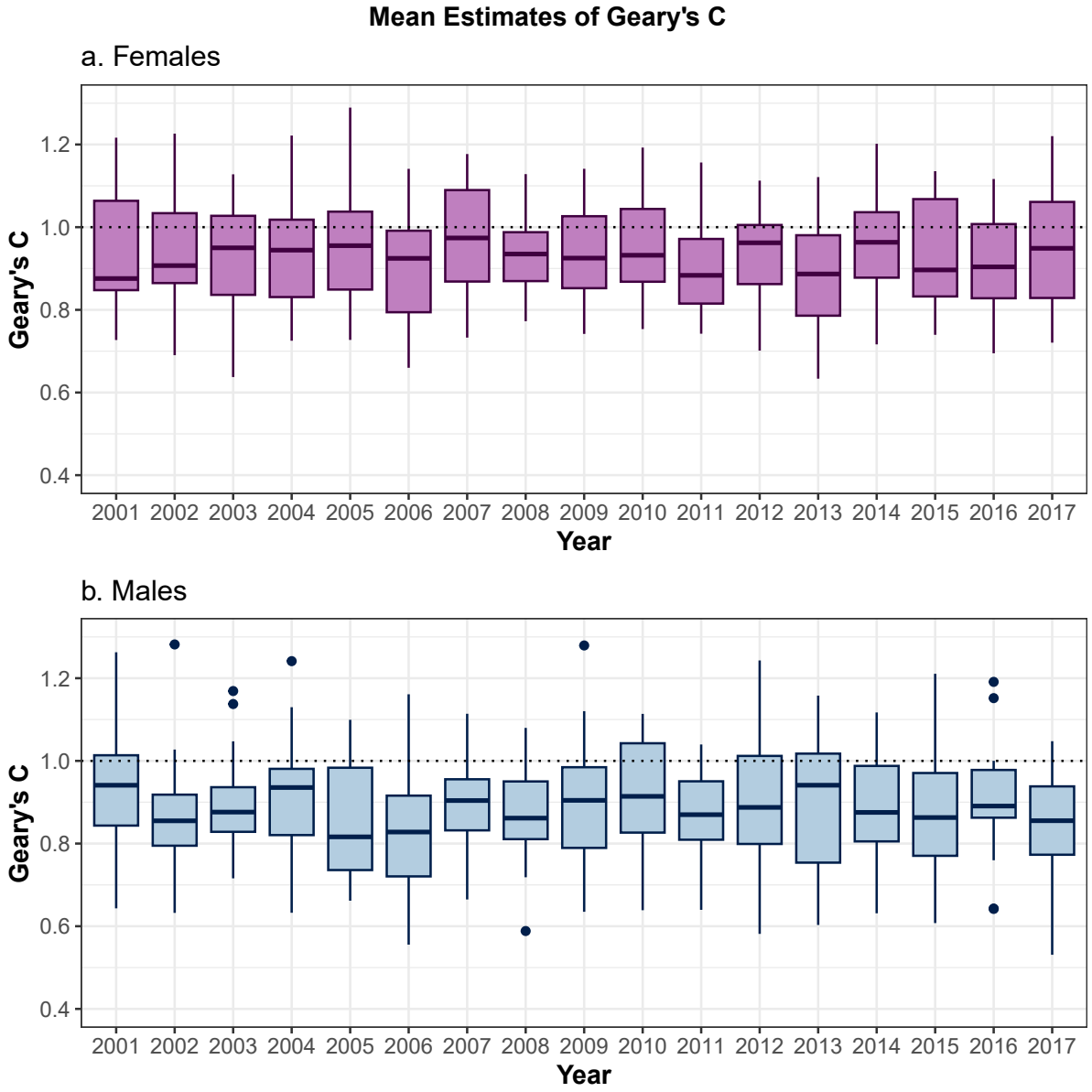


Figure A.2: The distribution of Geary's C values over all age groups by year. Estimates are given for both females (panel a) and males (panel b). The dashed line at 1 denotes the value of Geary's C under spatial independence.

Posterior Predictive Checks

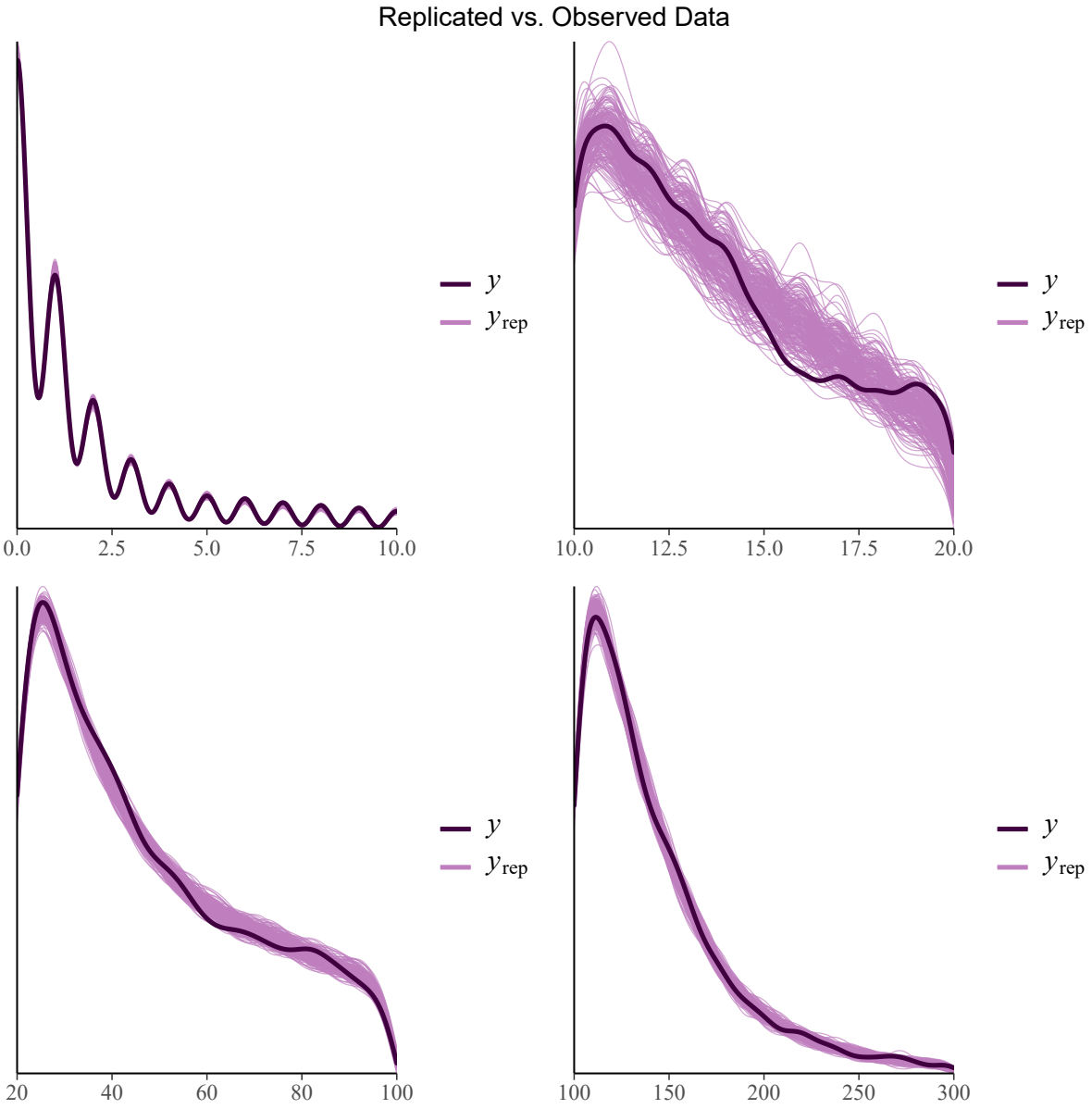


Figure A.3: Replicated vs. observed density of in-sample-values for females . The x-axis shows the actual value while the y-axis the respective density. Thick darker line denotes observed, while each light line denotes a realization of the replicated density. The plot is split along the x-axis for a better overview.

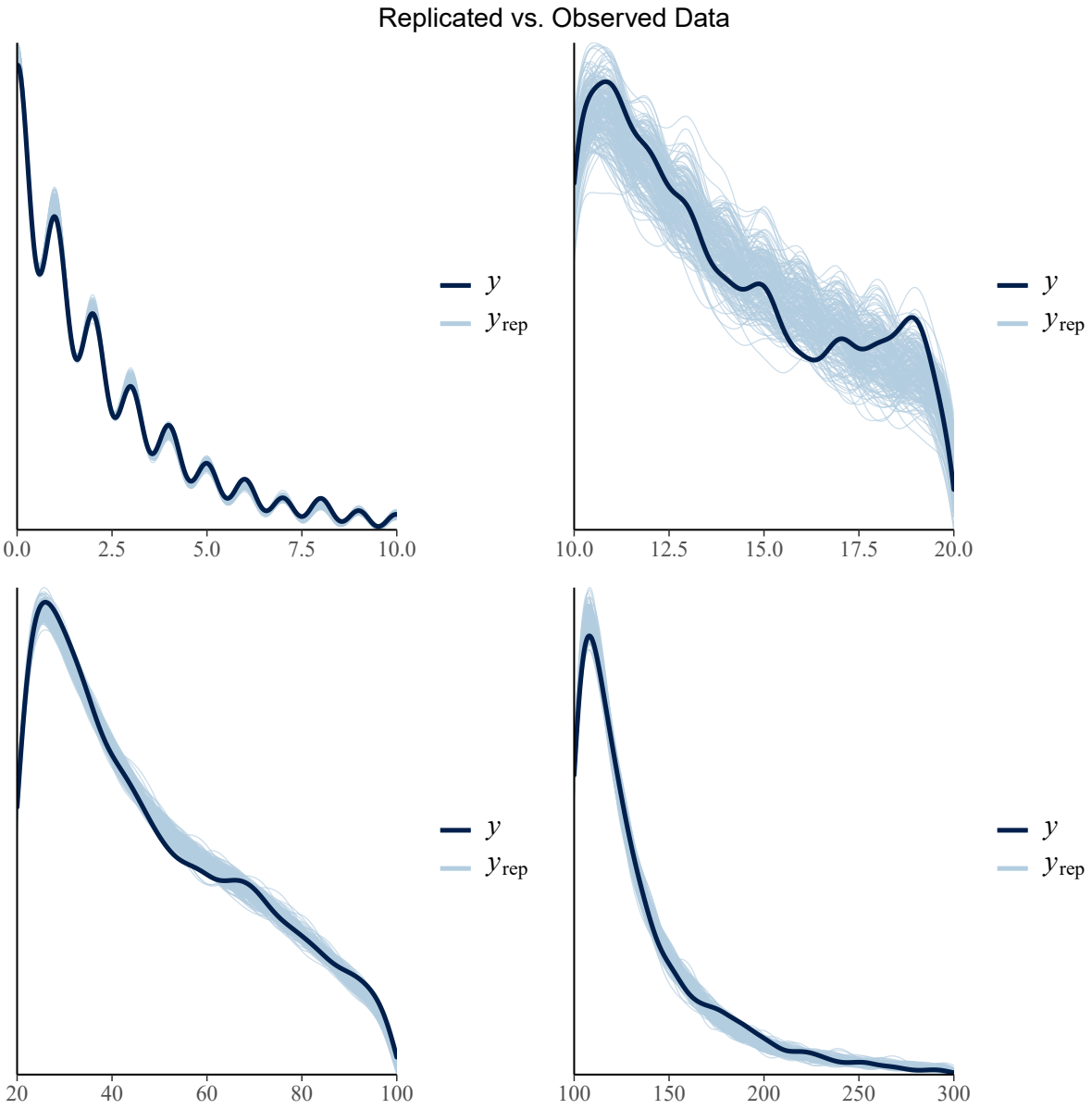


Figure A.4: Replicated vs. observed density of in-sample values for males. The x-axis shows the actual value while the y-axis the respective density. Thick darker line denotes observed, while each light line denotes a realization of the replicated density. The plot is split along the x-axis for a better overview.

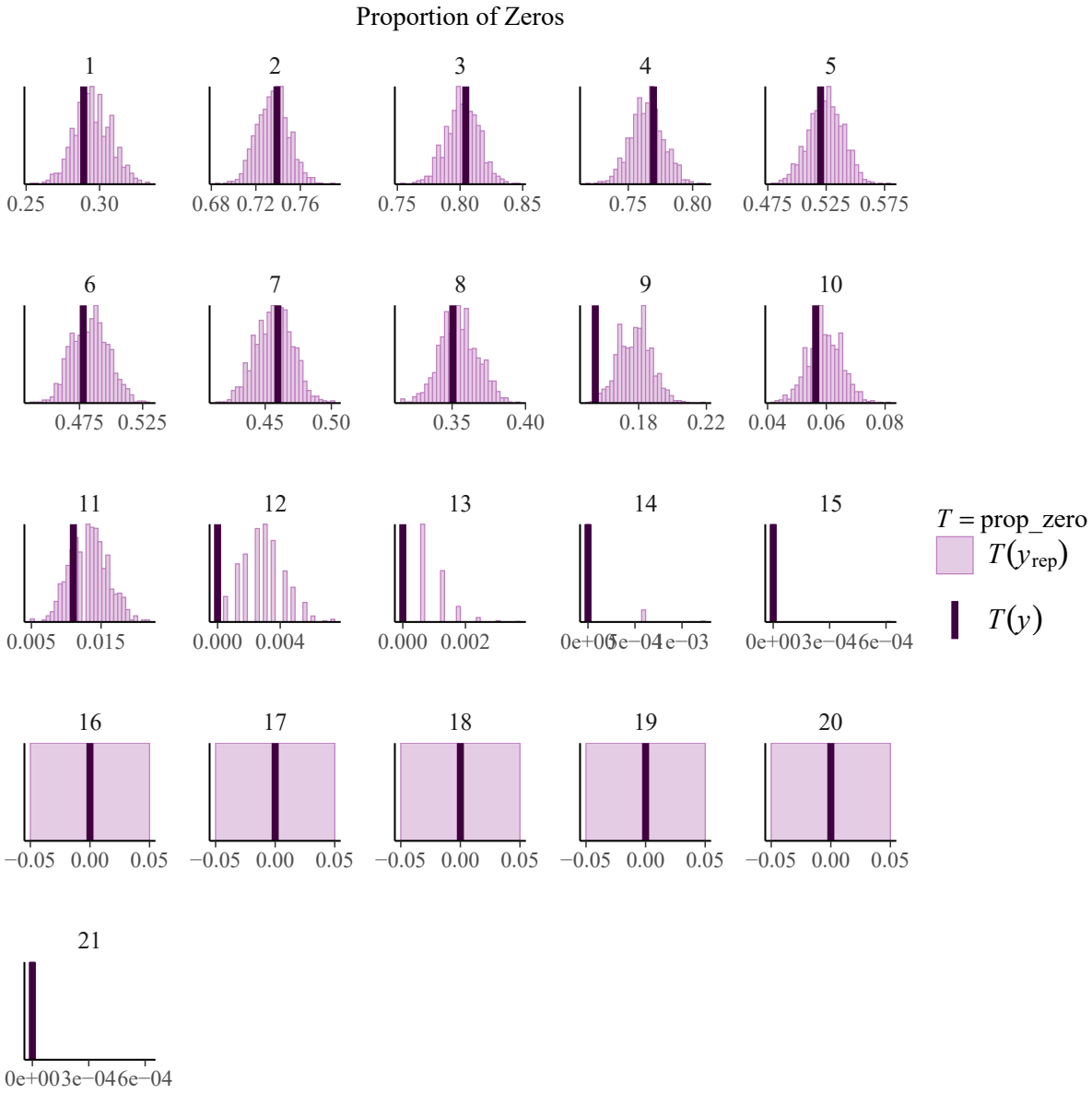


Figure A.5: Histogram of the test statistic $T[\boldsymbol{y}_{\boldsymbol{rep}}^{(\boldsymbol{x})}, \boldsymbol{\theta}]$ proportion of zeros, for the replicated data set. The observed data test statistic is given by the vertical line for all age groups of the APC_BYM2 for females.

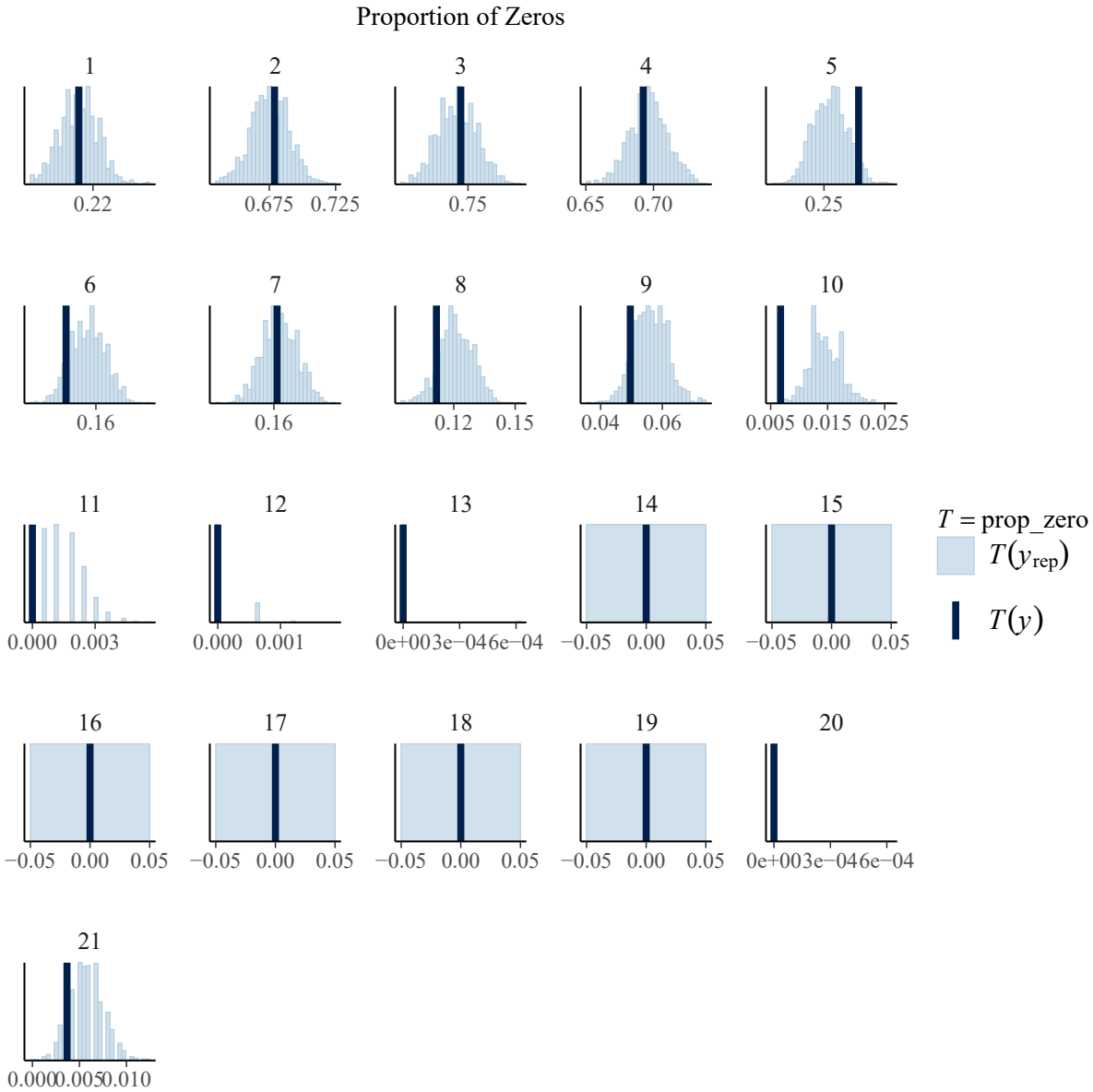


Figure A.6: Histogram of the test statistic $T[\boldsymbol{y}_{\boldsymbol{rep}}^{(\boldsymbol{x})}, \boldsymbol{\theta}]$ proportion of zeros, for the replicated data set. The observed data test statistic is given by the vertical line for all age groups of the APC_BYM2 for males.

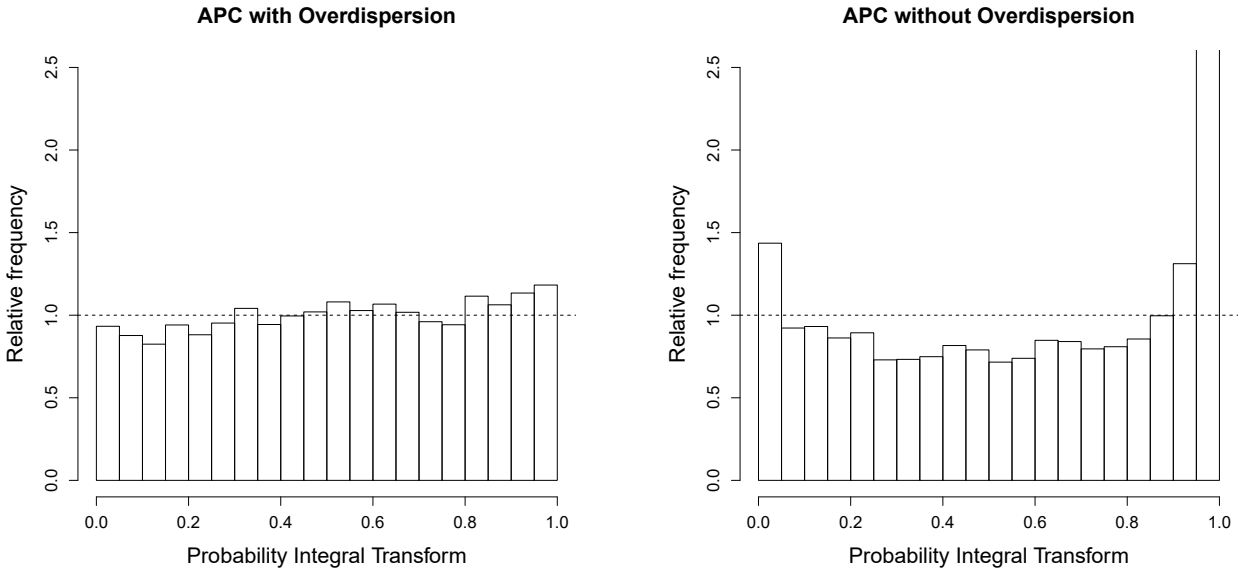


Figure A.7: Nonrandomized PIT for probabilistic forecast of APC Model for females with overdispersion parameter $\varepsilon_{x,t,r}$ (left) compared with probabilistic forecast of frequentistic APC Model without overdispersion parameter (right).

Estimated Life Expectancy

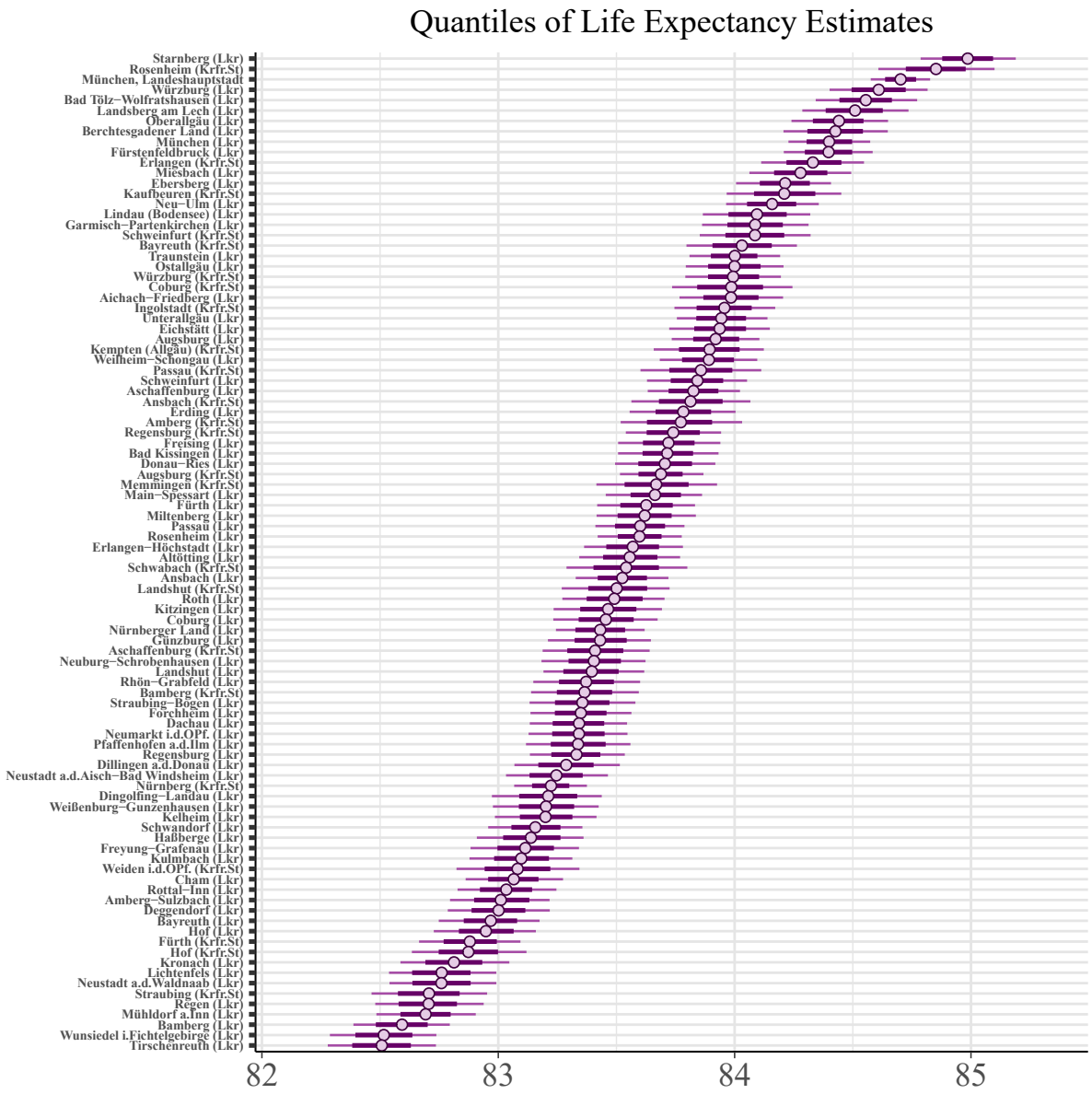


Figure A.8: Female life expectancy estimates in 2017. Points denote the mean value. The thick bar denotes the 50% while the thin the 80%- prediction interval

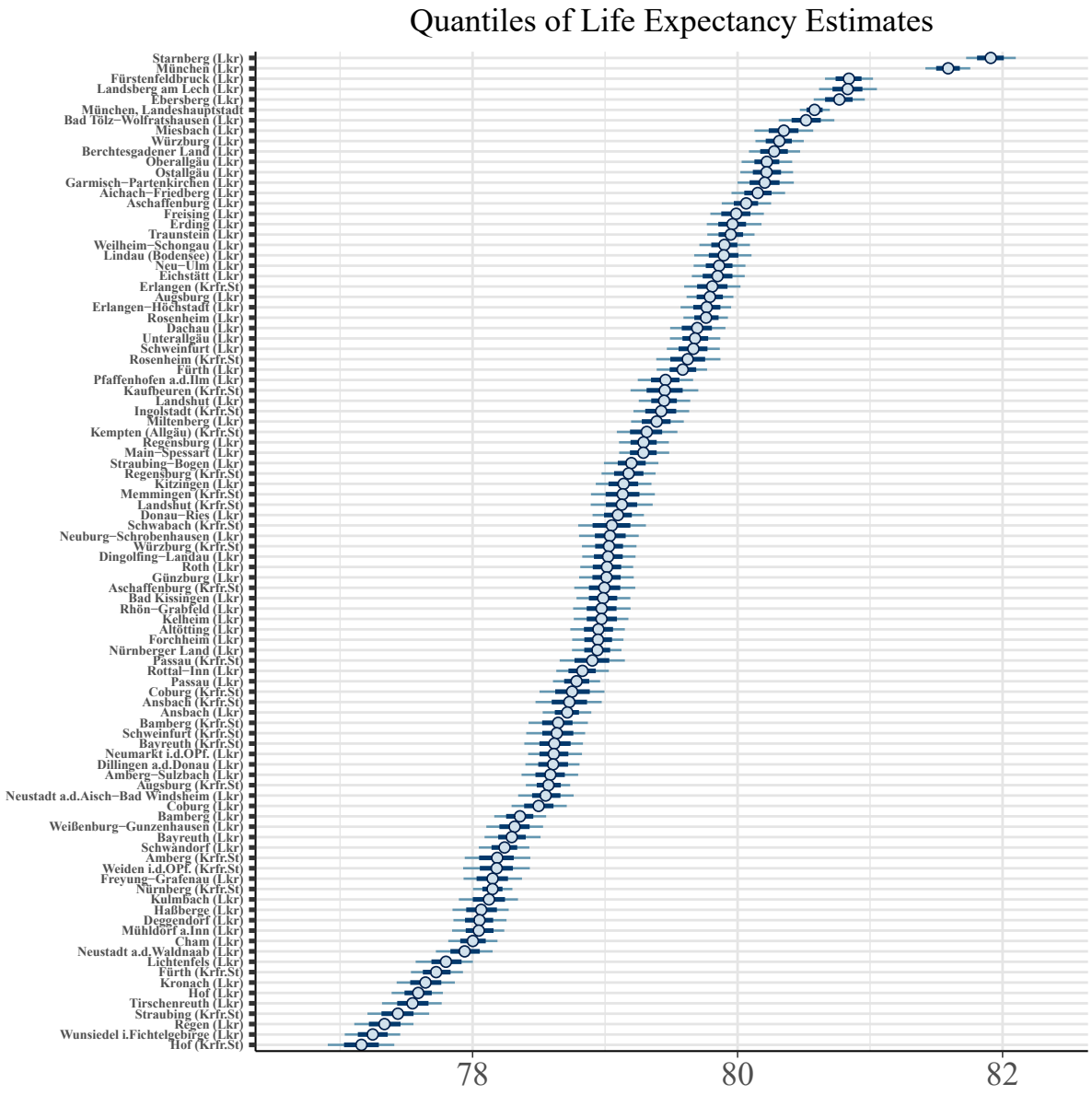


Figure A.9: Male life expectancy estimates in 2017. Points denote the mean value. The thick bar denotes the 50% while the thin the 80%- prediction interval

Chapter 2

Bayesian Mortality Modelling with Pandemics: A Vanishing Jump Approach[1]

Abstract

This paper extends the Lee-Carter model for single- and multi-populations to account for pandemic jump effects of vanishing kind, allowing for a more comprehensive and accurate representation of mortality rates during a pandemic, characterized by a high impact at the beginning and gradually vanishing effects over subsequent periods. While the Lee-Carter model is effective in capturing mortality trends, it may not always be able to account for large, unexpected jumps in mortality rates caused by pandemics or wars. Existing models allow either for transient jumps with an effect of one period only or persistent jumps. However, there is no literature on estimating mortality time series with jumps having an effect over a small number of periods, as is typically observed in pandemics. The Bayesian approach allows to quantify the uncertainty around the parameter estimates. Empirical data from the COVID-19 pandemic show the superiority of the proposed approach, compared to models with a transitory shock effect.

Keywords: Stochastic mortality modelling, Pandemic shocks, Jump effects, Bayesian inference.

2.1 Introduction

The model of Lee and Carter (1992) has been widely used in actuarial science and demography to forecast mortality rates based on past observations. This model assumes that mortality rates follow a stochastic trend with a time-dependent mortality factor, adjusted for age-specific effects, using two sets of age-dependent coefficients. While this model has been shown to be effective in capturing mortality trends in many countries, it

[1]Accepted version of article published in: Goes, J., Barigou, K., & Leucht, A. (2025). Bayesian mortality modelling with pandemics: A vanishing jump approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *74*(4), 1150–1182. https://doi.org/10.1093/jrsssc/qlaf018

may not be able to adequately account for large, unexpected jumps in mortality rates, such as those caused by pandemics (Chen et al., 2022; van Berkum et al., 2025). This paper aims to fill the gap by proposing a new modelling framework that is suitable for capturing the typical effects of a pandemic on the subsequent age-specific mortality rates. More specifically, we introduce single- and multi-population models that integrate serial dependence via vanishing shocks, where the shock's impact is highest in the beginning and then gradually diminishes in the subsequent years, allowing for a more reasonable description of the mortality rates during pandemics and a superior model fit.

We adopt a Bayesian framework for mortality modelling, a decision driven by the inherent advantages of this approach. Specifically, the Bayesian methodology integrates the estimation and forecasting processes, ensuring more consistent and robust estimates, as underscored by Cairns et al. (2011) and Wong et al. (2018). Furthermore, this approach excels in accommodating various sources of uncertainty in a coherent manner. Within the literature on Bayesian mortality modelling, Czado et al. (2005) pioneered the application of a comprehensive Bayesian approach specific to the Poisson Lee–Carter (LC) model. This methodology was later expanded to a multi-population context by Antonio et al. (2015). Other notable contributions include Pedroza (2006), who employed a Bayesian state-space model using Kalman filters to address missing data issues in mortality forecasting. The development and wider availability of Markov chain Monte Carlo (MCMC) techniques have further increased the application of Bayesian methodologies in mortality modelling, as evidenced by the growing body of recent work including (e.g. Alexopoulos et al., 2019; Barigou et al., 2023; Li et al., 2019; Wong et al., 2023).

The COVID-19 pandemic has emphasised the importance of incorporating mortality shocks into the LC model. While jump effects have been introduced in previous studies, they often fail to account for age-specific pandemic effects. Cox et al. (2006) and Chen and Cox (2009) proposed extensions with permanent and transitory jump effects, respectively, but the jump effect is applied to the time-dependent factor instead of the mortality rates. This means that the age pattern of a potential shock is identical to that of the general mortality improvement. To address this shortcoming, Liu and Li (2015) extended the Lee-Carter model by including a time- and age-dependent jump effect, which allows, for example, to capture the age-specific effect of COVID-19. However, their method only allows the inclusion of transitory mortality shocks that last one period, i.e. one year. Here, these effects are incorporated using independent and identically distributed (i.i.d.) shock variables. Consequently, during years of a pandemic, mortality rates experience an upward shift. However, this model has a limitation in its assumption of a time-independent jump variable. Specifically, it implies that during years with consecutive shocks, the severity of mortality rate adjustments is independent, or for years following a single-period shock, the effect completely vanishes in the subsequent year.

Such assumptions are inconsistent with observed pandemic patterns, wherein mortality rates are heavily impacted during the initial stages and gradually taper off.

In the aftermath of the COVID pandemic, a variety of models were proposed to capture the nuances of mortality trends. For example, van Berkum et al. (2025) extended the multi-population model originally proposed by Li and Lee (2005) for this purpose. Their model integrates three layers: the first two focusing on pre-COVID mortality trends and the third specifically capturing the excess mortality attributable to COVID. The framework offers mortality forecasts based on a spectrum of potential pandemic trajectories, determined by a parameter that varies between 0 and 1, yet remains uncalibrated. Meanwhile, Zhou and Li (2022) introduced a tri-level model to simulate future mortality scenarios influenced by events similar to the COVID outbreak. The intricacies of the pandemic's progression are encapsulated within their model's third layer, which is heavily informed by expert insights. Further broadening the scope, Chen et al. (2022) developed a multi-country mortality framework which incorporates two distinct jump components: one signifying global pandemic shocks, and the other reflecting country-specific disturbances. Robben and Antonio (2024) applied a multi-population regime switching model to switch between periods of high volatility states (i.e. shock years) and low volatility states. However, they also assume uncorrelated one-period shocks. Richards (2024), on the other hand, discusses techniques for robust estimation of mortality rates in the presence of outliers, with a scenario included for gradually diminishing effects.

All previous literature that attempts to account for vanishing jumps uses either expert opinion or simulation studies based on possible vanishing scenarios. Furthermore, the estimation procedure is typically based on a frequentist, multi-step process. To overcome these issues and better reflect the dynamic nature of a pandemic, we present a modelling framework that allows for the inclusion of serially dependent shock components. As an example, we propose two ways of modelling the serial dependence: an autoregressive structure and a moving average type structure. For both in-sample and out-of-sample data, we demonstrate that our models outperform those of Liu and Li (2015), where the shock components remain independent. In addition to single populations, we adapt our model to multi-populations as well, showing that this leads to an even better in-sample fit. Lastly, we prove parameter identification for both frameworks.

The remainder of this paper is organised as follows. Section 2.2 provides the specification of our single-population model, which is composed of a baseline Lee-Carter model and a vanishing jump component with age-specific effects, which can be autoregressive or moving average. Section 2.3 details the estimation procedure while section 2.5 introduces the methods that we apply for the comparison and evaluation of our methods. Section 2.6 studies in-sample performance based on COVID-19 data from the United States, Spain

and Poland. In particular, we compare the performance of our model to the original Lee-Carter model and to the model of Liu and Li (2015), which does not include gradually vanishing effects. Section 2.7 discusses out-of-sample performance based on world wars data for England and Wales, both during times of war and normal times. Section 2.8 proposes a multi-population extension with vanishing jumps and compares single and multi-population models. Finally, Section 2.9 provides concluding remarks.

2.2 Model specification

Our proposed model builds upon previous work on the Lee-Carter model and its extension with short-term jump effects by Liu and Li (2015). Our model can be seen as a generalisation of their approach, allowing for more flexibility and accuracy in capturing pandemic effects on mortality rates. In this section, we start with a brief overview of both the Lee-Carter model and the Liu-Li model to establish the foundation for our proposed model.

When studying human mortality, the data at hand consist of death counts $D_{x,t}$ and central exposures $E_{x,t}$, where $x \in \{1, 2, \ldots, A\}$ and $t \in \{1, 2, \ldots, T\}$ represent a set of A age groups and T calendar years, respectively. We denote by $m_{x,t}$ the central death rate at age x and calendar year t, given by

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}}.$$

2.2.1 The Lee-Carter model

The Lee-Carter model (Lee & Carter, 1992) is a well-known method for modelling mortality rates over time. It assumes that the logarithm of the central death rate $m_{x,t}$ for age group x in year t can be expressed as

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

where α_x represents the static level of mortality for age group x, κ_t captures the variation of log mortality rates over time, β_x measures the sensitivity of $\ln(m_{x,t})$ to changes in κ_t, and $e_{x,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$ is the error term. In the Lee-Carter model, α_x and β_x represent age-specific effects, while κ_t is a time-varying factor that captures the overall trend in mortality rates over time. Regarding the estimation, the frequentist approach is usually performed in two steps. First, parameters are obtained by maximising the model log-likelihood, and then in a second step, projections are made by time-series techniques (Pitacco, 2009). In a Bayesian approach, the estimation and forecasting steps are performed in a single step, ensuring more consistent estimates in the estimation procedure

(Cairns et al., 2011).

Despite its success in modelling mortality rates over time, the Lee-Carter model has a severe limitation when it comes to pandemics, as the model assumes that mortality rates evolve smoothly over time, without sudden changes or shocks, driven by a simple random walk with drift.

2.2.2 The Liu-Li model

To introduce short-term jump effects, Liu and Li (2015) proposed an extension of the original Lee-Carter model, which includes an extra jump term as follows:

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t + N_t J_{x,t} + e_{x,t}.$$

Here, α_x, β_x and $e_{x,t}$ have the same meanings as in the original Lee-Carter model, while κ_t is assumed to be a random walk with drift. Additionally, N_t represents a binary random variable that equals one if a mortality jump occurs in year t and zero otherwise. The authors assume that the N_t's are i.i.d. Bernoulli distributed with parameter p, denoting the probability of a mortality jump in a calendar year. $J_{x,t}$ measures the effect of a mortality jump that occurred in year t on age group x.

Three specific model variants were proposed, denoted as models J0-J1-J2. Model J1 is the closest to our model and is given by

$$\ln(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \beta_x^{(J)} N_t Y_t + e_{x,t}, \tag{2.1}$$

where Y_t denotes the effect or severity of the mortality jump at time t. These jump effects are assumed to be i.i.d. Gaussian variables. Compared to the Lee-Carter model, a new age pattern of pandemic effects $\beta_x^{(J)}$ is introduced. It is multiplied by the pandemic jump effect to capture age-specific variation, that is different from the period effects β_x.

The model proposed by Liu and Li provides a valuable extension of the Lee-Carter model as it allows the inclusion of short jumps and different age patterns. However, it has two weaknesses: first, this model assumes that age patterns of different mortality shocks are the same, while historically, different age patterns have been observed. For example, half of the deaths caused by the 1918 flu pandemic occurred among 20- to 40-year-olds (Gagnon et al., 2013), while COVID-19 has affected mostly the most vulnerable people (Ferguson et al., 2020; O'Driscoll et al., 2021). In a Bayesian setting, however, the pandemic effects $\beta_x^{(J)}$ are not considered to be fixed. The use of different priors provides a wide range of estimation possibilities. Second, the yearly jumps are independent, and the Liu-Li model does not allow for a jump event lasting over several years with a vanishing effect, as it can be observed for COVID-19. Our model, presented in the next section,

addresses these shortcomings of the Liu-Li model.

2.2.3 A new class of models allowing for serial dependent jump effects

The limitations of the Lee-Carter model and its extension by Liu and Li (2015) motivate the need for a more flexible model that can capture the influence of a pandemic lasting over several years with a vanishing effect. Several recent studies have addressed this issue by proposing extensions to the Lee-Carter model. For instance, van Berkum et al. (2025) and Zhou and Li (2022) discussed the idea of a vanishing effect in the context of COVID-19, but did not try to estimate a corresponding parameter. To address this, we propose a model that allows for pandemic shocks that are transitory and vanishing over time. Our model, formulated in (2.2), extends the Lee-Carter model by adding a pandemic shock component, J_t, that captures serial dependence.

The base line model is very similar to that of Liu and Li (2015) and given by

$$\ln\left(m_{x,t}\right) = \alpha_x + \beta_x \kappa_t + \beta_x^{(J)} J_t + e_{x,t} \tag{2.2}$$

Like Liu and Li (2015), we model the time effect using a random walk representation

$$\kappa_t = \kappa_{t-1} + d + \xi_t, \tag{2.3}$$

where $d \neq 0$ denotes the drift parameter and $(\xi_t)_t$ is a sequence of error terms. However, instead of setting $J_t = N_t\, Y_t$ as in Section 2.2.2, where Y_t denotes the magnitude of the jump effect and $N_t \in \{0, 1\}$ indicates the jump occurrences, we introduce a more flexible approach that allows for a single shock to have an effect on consecutive years using ideas from time series models. We propose two options to capture the serial dependence of the jump parameter J_t, namely an autoregressive structure in Subsection 2.2.3 and a moving average structure in Subsection 2.2.3. We will refer to the former as the AR model and the latter as the MA model from hereafter. Of course, other structures to model serial dependence are possible. Both of our models have the advantage of modelling more complex patterns of mortality shocks and are easy to interpret. While the Liu-Li model only provides an estimate on the severity of a mortality shock, our model adaptation provides an estimate on the severity, including how quickly the observed shock vanishes. This can help policymakers determine how long the effect of a particular mortality shock has lasted, and potentially adjust responses to future shocks.

Additionally, the Bayesian framework we use for parameter estimation provides a full predictive distribution for all parameters, including $\beta_x^{(J)}$, which enables us to account for uncertainty. This flexibility allows us, for example, to accommodate different age patterns for different mortality jumps, that is, for each future shock, we can obtain

different realisations of the age patterns, which was one of the limitations of the Liu-Li model.

Autoregressive structure

The discussions by van Berkum et al. (2025) and Zhou and Li (2022) suggest the use of an autoregressive type structure, where a pandemic effect slowly vanishes over time controlled by some parameter $a \in [0, 1)$. This is also in line with Markovian epidemiological models for health states of COVID patients, see e. g. Bartolucci et al. (2021). More precisely, we propose the following model:

$$\begin{aligned} \ln(m_{x,t}) &= \alpha_x + \beta_x \kappa_t + \beta_x^{(J)} J_t + e_{x,t}, \\ J_t &= a\, J_{t-1} + N_t\, Y_t. \end{aligned} \tag{2.4}$$

The parameters Y_t and N_t have the same meaning as in the model of Liu and Li (2015). However, the jump size J_t, which captures the impact of the pandemic shock on mortality rates, includes a vanishing effect controlled by the parameter $a \in [0, 1)$. The model (2.4) is a (Bayesian) generalisation of the model J1 in Equation (2.1) of Liu and Li (2015), which is recovered when $a = 0$ and allows for a gradually vanishing effect when $a > 0$. Indeed, if there is a mortality jump in year t, the impact on the log mortality rates is given by $\beta_x^{(J)} Y_t$ in year t, by $a\, \beta_x^{(J)} Y_t$ in year $t + 1$, and so on.

Moving average structure

Instead of allowing the shock effect to slowly vanish over time, we may also assume a moving average type structure with order Q. Here, the initial shock has an effect for a total of Q periods and then disappears completely. A reasonable order Q, which indicates how long the initial shock lasts in the subsequent periods, must be selected manually, a potentially challenging task. In addition, as the order Q increases, the number of parameters increases as well, making the identification and estimation of parameters more complex. For the sake of simplicity, we therefore assume a MA(1) structure which can be defined as

$$\begin{aligned} \ln(m_{x,t}) &= \alpha_x + \beta_x \kappa_t + \beta_x^{(J)} J_t + e_{x,t}, \\ J_t &= N_t\, Y_t + b\, N_{t-1}\, Y_{t-1}. \end{aligned} \tag{2.5}$$

Similar to the AR model, we assume that $b \in [0, 1)$. Hence, the MA(1) model of (2.4) can be seen as a generalisation of the Liu and Li (2015) model that is recovered if $b = 0$.

2.3 Estimation procedure

For the estimation of mortality models, there are two common routes: either estimate the model on the central death rates directly (the traditional approach), or estimate the model on the first differences of the log mortality rates, also called mortality improvements. These two variants are referred to as *Route I* and *Route II* in the terminology of Haberman and Renshaw (2012). In this paper, we proceed to Route II. It has the advantage of eliminating the static age effect α_x of the model, thereby reducing the number of identifiability constraints needed (see Hunt and Blake (2020) and the next Section 2.3.1). We remark that Mitchell et al. (2013) conducted an extensive study comparing the Route I and Route II approach of multiple models in terms of in-sample fit, including the LC model as well as variants thereof, and found the Route II method to be superior. Moreover, Wong et al. (2023) compared the in-sample fit of a LC model estimated on mortality rates with that of an age-period model estimated on mortality improvement rates and found the latter to be superior. In addition, Mitchell et al. (2013) compared the width of the prediction intervals on held out data and found the ones of the Route II approach to be narrower and more accurate in terms of coverage.

In the Route II approach, the mortality improvements are modelled directly, defined as

$$Z_{x,t} \coloneqq \ln\left(m_{x,t+1}\right) - \ln\left(m_{x,t}\right). \tag{2.6}$$

Positive values of mortality improvement rates indicate worsening mortality conditions relative to the previous year, while negative mortality improvement rates indicate an improvement in mortality. Our model specification in (2.2) can then be written as

$$Z_{x,t} = \beta_x \left(\kappa_{t+1} - \kappa_t\right) + \beta_x^{(J)} \left(J_{t+1} - J_t\right) + \varepsilon_{x,t},$$

where $\varepsilon_{x,t} = e_{x,t+1} - e_{x,t}$. It follows, that $\varepsilon_{x,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_r^2)$ with $\sigma_r^2 = 2\sigma_e^2$. Using (2.3), this can be restated as

$$\begin{aligned} Z_{x,t} &= \beta_x \, \Delta\kappa_{t+1} + \beta_x^{(J)} \, \Delta J_{t+1} + \varepsilon_{x,t} \\ &= \beta_x \left(d + \xi_{t+1}\right) + \beta_x^{(J)} \, \Delta J_{t+1} + \varepsilon_{x,t}, \end{aligned} \tag{2.7}$$

where $\Delta J_{t+1} = J_{t+1} - J_t$ and, similarly, $\Delta\kappa_{t+1} = \kappa_{t+1} - \kappa_t$.

2.3.1 Identifiability constraints

Several mortality models, including the LC model, suffer from non-identifiability issues, meaning that different sets of parameters result in equivalent likelihoods and consequently the same fitted rates. In their paper, Hunt and Blake (2020) discuss the problem of non-

identifiability in LC type models with multiple age or period functions at length and provide a general theorem for the selection of suitable constraints. Since our proposed model can be seen as an extension thereof, their logic can be applied to find the number of needed constraints.

In a standard LC model, the age effect may be scaled and the time effect shifted to produce a new set of parameters resulting in the same fitted mortality rates. Thus, the parameters are not uniquely determined and can be transformed in two ways, namely

$$\{\tilde{\alpha}_x,\, \tilde{\beta}_x,\, \tilde{\kappa}_t\} = \left\{\alpha_x,\, \frac{1}{a}\,\beta_x,\, a\,\kappa_t\right\}, \tag{2.8}$$

$$\{\tilde{\alpha}_x,\, \tilde{\beta}_x,\, \tilde{\kappa}_t\} = \{\alpha_x - b\,\beta_x,\, \beta_x,\, \kappa_t + b\}, \tag{2.9}$$

for all $x \in \{1, 2, \ldots, A\}$ and $t \in \{1, 2, \ldots, T\}$.

In principle, the same problem holds for our model formulation as well. Using matrix notation we can rewrite the first equation of (2.2) in a compact way. Let $\boldsymbol{B}_x = \left(\beta_x, \beta_x^{(J)}\right)^\top$ and $\boldsymbol{K}_t = \left(\kappa_t, J_t\right)^\top$, then

$$\ln(m_{x,t}) = \alpha_x + \boldsymbol{B}_x^\top \boldsymbol{K}_t + e_{x,t}. \tag{2.10}$$

The model in (2.10) has the same structure as the classical LC model and can be thought of as a multivariate extension, coined LC2 in the terminology of Hunt and Blake (2020). Unsurprisingly, the model in (2.10) suffers from non-identifiability. Let there be a matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ that is invertible and a matrix $\mathbf{D} \in \mathbb{R}^{2\times 1}$. Then, according to Hunt and Blake (2020, Theorem 1), equations (2.8) and (2.9) can be generalised to higher dimensions where the parameters of (2.10) can be transformed using

$$\{\tilde{\alpha}, \tilde{\boldsymbol{B}}_x, \tilde{\boldsymbol{K}}_t\} = \left\{\alpha_x, \mathbf{A}^{-1}\boldsymbol{B}_x, \mathbf{A}\boldsymbol{K}_t\right\} \tag{2.11}$$

$$\{\tilde{\alpha}, \tilde{\boldsymbol{B}}_x, \tilde{\boldsymbol{K}}_t\} = \left\{\alpha_x - \mathbf{D}^\top \boldsymbol{B}_x, \boldsymbol{B}_x, \boldsymbol{K}_t + \mathbf{D}\right\}. \tag{2.12}$$

Since matrix $\mathbf{A}$ is (2×2) and $\mathbf{D}$ is (2×1), there are in total six free parameters meaning that we have to impose six identifiability constraints for the model in (2.10).

However, note that Equation (2.10) includes an age specific intercept. When differencing the log mortality rates to obtain $Z_{x,t}$, i.e. applying the Route II estimation method, the age specific intercept cancels. In this case, we obtain that $\mathbf{D} = 0_{2\times 1}$ in (2.12) and no further identifiability issues arise from (2.12). Consequently, there is a reduced set of identifiability constraints, namely the four entries of the matrix $\mathbf{A}$, as only transfor-

mations of (2.11) are relevant (Hunt & Blake, 2020, Appendix A.). Hence, by applying the Route II estimation approach we can reduce the amount of identifiability constraints needed from six to four, by cancelling out the static age function α_x due to differentiation of the death rates.

As we prove in the Appendix, identification is ensured by imposing the standard sum-to-one constraints on age parameters, that is

$$\sum_{x=1}^{A} \beta_x = 1 \quad \text{and} \quad \sum_{x=1}^{A} \beta_x^{(J)} = 1,$$

and by using corner constraints on the first differenced time parameters, $\Delta J_2 = 0$ and $\xi_2 = 0$, resulting in a total of four constraints. This is enough to identify the drift d and the parameters of the mortality improvement rates $Z_{x,t}$, i.e. β_x, $\beta_x^{(J)}$, $\Delta\kappa_t$, and ΔJ_t. If we assume additionally that $J_1 = J_2 = 0$, then we can identify all of $(J_t)_t$ iteratively using $(\Delta J_t)_t$.

However, depending on the model structure assumed for J_t, additional constraints need to be imposed to identify the jump occurrences N_t and either the autoregressive parameter a or the moving average parameter b. In particular, we assume knowledge of a known time point $\tilde{t}$ after a first shock event where there is no jump, that is, $N_{\tilde{t}} = 0$. For example, this can be set to $\tilde{t} = T$. For the MA setting, some further constraints are needed. Details can be found in the Appendix B.1.

It should be noted that these are not the only identification constraints that can be set. Given the recommendation by Hunt and Blake (2020), another possibility is to adopt a true normalisation scheme for the age parameters. That is, instead of a sum-to-one constraint, we could set the age parameters to have an Euclidean norm of one, that is

$$\|\beta_x\|_2^2 = \sum_{x=1}^{A} (\beta_x)^2 = 1 \quad \text{and} \quad \left\|\beta_x^{(J)}\right\|_2^2 = \sum_{x=1}^{A} \left(\beta_x^{(J)}\right)^2 = 1.$$

The above identification can be achieved, for example, using **QR** decomposition, which results in two orthonormal age vectors that do not only have a norm of one, but are also orthogonal to each other, resulting in a dot product of zero, that is $\sum_x \beta_x \beta_x^{(J)} = 0$. When adopting the identification scheme using **QR**, the corner constraint on the time dependent parameters needs only to be set on the jump effect, thus $\Delta J_2 = 0$.

Both identification schemes have been successfully implemented and give unique parameter estimates. For sake of model comparison, we choose to go with the standard sum-to-one constraints, as these are the ones selected by Liu and Li (2015).

2.3.2 Priors

To estimate the parameters in Equation (2.7), we consider a Bayesian approach to inference. It is based on the idea of updating prior beliefs with the data at hand to obtain a posterior distribution of the parameters. For the selection of priors there are many options available. If not stated otherwise, we employ the use of so-called weakly informative priors. Here, the prior should rule out unreasonable values but not be too restrictive so that it rules out plausible values.

As stated in Equation (2.3), we assume that the time-dependent parameter κ_t follows a random walk type representation, which we model using a normal prior: $\Delta\kappa_t \overset{i.i.d.}{\sim} \mathcal{N}(d, \xi_t)$. The jump effects are given a half normal prior, such that, $Y_t \overset{i.i.d.}{\sim} \mathcal{N}^+(\mu_Y, \sigma_Y)$, because the focus on this paper is on catastrophic mortality jumps. This is in contrast to the approach of Liu and Li (2015), where Y_t is assumed to be Gaussian to preserve the tractability of the likelihood. However, they note that other distributions might produce a better fit. The jump occurrence is modelled using a Bernoulli distribution $N_t \overset{i.i.d.}{\sim} \text{Bern}(p)$. The age-specific parameters $(\beta_1, \ldots, \beta_A)$ and $\left(\beta_1^{(J)}, \ldots, \beta_A^{(J)}\right)$ are given multivariate Dirichlet priors, which implicitly impose the sum-to-one constraints. For both the autoregressive parameter a as well as the moving average parameter b we assume a slightly informative normal prior truncated from zero to one with a mean of zero and a standard deviation of 0.4. This parameterisation favours smaller values of a respectively b with the most prior mass around zero. Hence, there needs to be evidence by the likelihood to move the posterior estimate away from zero.

For the hyperparameters we choose a mix between weakly informative and informative priors. Starting with the drift parameter d we assume a normal prior with a smaller standard deviation. The normal prior guarantees that $\Pr(d \neq 0) = 1$, as required for identification of our parameters. For the jump probability we impose a rather informative hyperprior, where $p \sim \text{Beta}(1, 20)$, which strongly favours small values of p. This is for the following reason. The parameter J_t is intended to model extreme events, not just noise. Thus, a shock should be something that occurs rarely, less than every few years, rather than some regular ups and downs. The latter type of effects should be captured by the noise term of the random walk coefficient and not considered a shock. Thus we know in advance that mortality shocks appear infrequently and that the jump proportion has to be low. This information can be included in the prior. Moreover, when experimenting with uninformative hyperprior settings, e.g. the well-known Jefferys prior $p \sim \text{Beta}(0.5, 0.5)$, we noticed the tendency of the model to flag many more smaller bumps as potential shocks which sometimes led to a lack of convergence of the age-specific jump parameters $\beta_x^{(J)}$. This effect can be explained by the fact that our data contains too few shock events to calibrate reasonably well without any prior information. Moreover, several papers in

the literature rely on expert knowledge to handle this issue. Imposing an informative prior on p alleviated the problem in a data-driven way based on the comparatively mild information of low shock frequency and helped with convergence. However, we note that our specific choice of p is of course subjective. We could have used a different parameterisation, e.g. $p \sim \text{Beta}(1, 10)$ or $p \sim \text{Beta}(1, 15)$. A small sensitivity analysis showed that either hyperprior led to a similar posterior.

Using our modelling approach, we want to capture shocks that have a positive (i.e. increasing) effect on death rates. We therefore assume that the jump effect Y_t can only take on positive values and impose a half normal prior, which assumes a positive mean. Therefore, we choose a half-normal prior on the jump mean parameter μ_Y as well. Lastly, all standard deviations are given half-normal priors. An overview on specific values for the hyperparameters can be found in the Appendix B.4.

2.4 Parameter estimation

To estimate the parameters of our proposed model, we use `NIMBLE` (de Valpine et al., 2017), a system for programming statistical algorithms in `R`. `NIMBLE` provides a flexible and intuitive framework for model specification while supporting programming functions that adapt to model structures. Moreover, it allows for the selection of multiple samplers that include the well known MCMC methods as well as Hamiltonian Monte Carlo (HMC). For each parameter a different sampler can be chosen allowing for great flexibility and efficient computation. `NIMBLE` can be accessed via the `nimble` package in R (de Valpine et al., 2024). To be able to use the HMC sampler, the package `nimbleHMC` (Turek et al., 2024) must be downloaded as well. If not stated otherwise, `NIMBLE` uses a conjugate Gibbs sampler where possible as well as Metropolis-Hastings. However, the latter tends to be very inefficient due to the high autocorrelation of the samples. As a result, we choose to change the samplers of multiple variables resulting in improved mixing performance. Moreover, to be able to use the flexibility of `NIMBLE`, we have chosen to parameterise the Dirichlet in terms of a normalised Gamma distribution, which allows us to select from a greater pool of available samplers, like the multivariate slice sampler of Tibbits et al. (2014) for example or an HMC sampler. A justification of the construction of a Dirichlet using the Gamma distribution can be found in the Appendix B.2). In addition, for the jump occurrence N_t, `NIMBLE` uses a special Gibbs sampler for binary-valued variables. A list of the specific samplers for each parameter can be found in the Appendix B.5.

To assess the convergence of all model parameters, we employ three widely recognised diagnostics: the split-$\hat{R}$ statistic as well as variants of the effective sample size, namely bulk effective sample size (Bulk-ESS) and tail effective sample size (Tail-ESS). These diagnostics are implemented using the `rstan` package (Stan Development Team, 2024a),

and for a more comprehensive discussion on their use, we refer to Vehtari et al. (2021).

In our analysis, we utilise two chains for each country and model. We discard the initial 7,500 iterations of each chain as "burn-in", ensuring that the chains have stabilised. Subsequently, we draw an additional 10,000 samples per chain, with only every 10th sample being retained for inference. For a comprehensive understanding of parameter convergence and additional details regarding the MCMC settings, please refer to the Appendix. Even without parallelisation, the total run time of the model is rather short and took around 5 minutes on an Intel i5-8365U CPU @ 1.60GHz processor with 16,0 GB RAM.

2.5 Model comparison

After having estimated the parameters we want to assess both the in as well as out-of-sample fit of our models.

2.5.1 In-sample comparison

To assess the in-sample fit of the models in question, we calculate the widely applicable or Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2010). A lower WAIC value indicates a better-fitting model among the alternatives, as is the case with most information criteria. What distinguishes WAIC is its fully Bayesian nature, since it considers the entire posterior distribution for model evaluation. In addition to WAIC, the in-sample fit may be compared using cross-validation.

Consider some data, $\mathbf{y} = (y_1, \ldots, y_N)$, which is modelled as independent given the parameter θ, hence $p(\mathbf{y}|\theta) = \prod_{i=1}^{N} p(y_i|\theta)$. A standard quantity in Bayesian analysis is the log predictive density (lpd),

$$\text{lpd} = \sum_{i=1}^{N} \log p(y_i|\mathbf{y}) = \sum_{i=1}^{N} \log \int p(y_i|\theta) p(\theta|\mathbf{y}) d\theta.$$

Using the lpd, we can calculate the WAIC by

$$\begin{aligned} \text{WAIC} &= -2\text{lpd} + 2p_{waic} \\ p_{waic} &= \sum_{i=1}^{N} \text{Var}\big(\log p(y_i|\theta)\big). \end{aligned}$$

To calculate the WAIC in practice, the lpd and p_{waic} have to be estimated using posterior draws. Let $\theta^{(s)}$ denote the s-th sample from the posterior, with $s = 1, \ldots, S$. The inner

expectation of the lpd can be approximated by:

$$\widehat{\text{lpd}} = \sum_{i=1}^{N} \log \left(\frac{1}{S} \sum_{s=1}^{S} p\left(y_i|\theta^{(s)}\right) \right).$$

An estimate of p_{waic} is given by

$$\widehat{p}_{waic} = \sum_{l=1}^{n} V_{s=1}^{S} \log p\left(y_l|\theta^{(s)}\right),$$

where $V_{s=1}^{S}$ represents the sample variance (Vehtari et al., 2017).

The Bayesian leave-one-out cross validation (LOO-CV) is based on the log predictive density given the data without the i-th data point $p(y_i|y_{-i})$. In practice, it is calculated as

$$\text{lpd}_{\text{loo}} = \sum_{i=1}^{N} \log p(y_i|y_{-i}),$$

where

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta$$

denotes the leave-one-out predictive density without the i-th data point. In the above setting, exact cross validation would require refitting the model N times. However, $p(y_i|y_{-i})$ can be approximated using importance sampling (Vehtari et al., 2021). A model with a higher lpd_{loo} indicates a better model fit by superior predictive performance. Oftentimes, lpd_{loo} is provided on the deviance scale, that is LOO-CV $= -2\text{lpd}_{\text{loo}}$, where a lower score suggests the better fit. Both the calculation of the WAIC and LOO-CV is implemented in the `loo` package in R (Vehtari et al., 2023).

2.5.2 Mortality forecasts

As our model estimates mortality improvement rates, the generated forecasts are of the same form. However, to obtain forecasts of future (log) death rates, the mortality improvement rates can be transformed. Using Equation (2.6), we can calculate the death rate at time $t+1$ as follows:

$$\ln\left(m_{x,t+1}\right) = \ln\left(m_{x,t}\right) + Z_{x,t}. \tag{2.13}$$

Let the projection periods be denoted as $h \in \{1, \ldots, H\}$. To generate a h-step ahead forecast of $\ln\left(\hat{m}_{x,T+h}\right)$, we proceed by generating new values from our posterior predictive distribution for $Z_{x,T+1}, \ldots, Z_{x,T+H}$ and apply Equation (2.13) recursively. This procedure can be repeated S times to obtain S draws from the posterior predictive distributions of $\ln\left(\hat{m}_{x,T+h}\right)$. We can then derive prediction intervals using Monte Carlo simulations.

To make predictions for future values of $Z_{x,T+1}, \ldots, Z_{x,T+H}$, we must also generate new values for time-dependent parameters. For example, to predict $Z_{x,T+1}$ for the model with the autoregressive structure of (2.4), we can follow these steps for each posterior draw ($s = 1, \ldots, S$):

Step 1: Generate new values of $N_{T+1}^{(s)}$ by first drawing a value of $p^{(s)}$ from the posterior distribution and then sampling $N_{T+1}^{(s)}$ from a Bernoulli distribution with parameter $p^{(s)}$.

Step 2: Generate new values of $J_{T+1}^{(s)}$. Start by drawing $\mu_Y^{(s)}$ and $\sigma_Y^{(s)}$ from the posterior distribution. Then sample a new value of $Y_{T+1}^{(s)}$ from a normal distribution with mean $\mu_Y^{(s)}$ and standard deviation $\sigma_Y^{(s)}$. Afterwards, draw $a^{(s)}$ and $J_T^{(s)}$ from the posterior distribution. Use the newly generated $N_{T+1}^{(s)}$ from Step 1 to compute a future value of $J_{T+1}^{(s)}$.

Step 3: Generate a new error term $\varepsilon_{x,T+1}^{(s)}$ by sampling from a normal distribution with mean 0 and standard deviation $\sigma_r^{(s)}$.

Step 4: Obtain the s-th posterior draw for the remaining parameters and substitute all values into Equation (2.7) to generate $Z_{x,T+1}^{(s)}$.

These steps are then repeated for $T+1, \ldots, T+H$ to generate future log death rates: $\ln(m_{x,t+1}), \ldots, \ln(m_{x,t+H})$.

2.5.3 Out-of-sample comparison

In addition to the in-sample comparison we can compare the predictive accuracy of our models on out-of-sample data. Since we are estimating the parameters in a Bayesian setting, our forecasts denote an entire predictive distribution rather than a single point, i.e. a mean forecast. Scoring rules provide a means to compare the accuracy of a predictive distribution of competing models around an observed data point. An overview on the idea as well as examples of scoring rules can be found in Gneiting and Raftery (2007). Scoring rules are similar to information criteria in the sense that a lower score denotes a better fit. Two popular examples of scoring rules are the negative log score and the continuous ranked probability score (CRPS), which is more robust.

Suppose that the data for T years is split into training and validation set, where $Z_{1:M}$ denotes the mortality improvement rates for all ages of the first M years used to fit the model with corresponding parameters $\boldsymbol{\theta}_M$. The log score (LogS) is given by the negative logarithm of the predictive density for a future h-step ahead observation and defined as

$$\text{LogS}(Z_{x,M+h}) = -\log p(Z_{x,M+h}|Z_{1:M}) = -\log \int p(Z_{x,M+h}|\boldsymbol{\theta}_M) p(\boldsymbol{\theta}_M|Z_{1:M}) d\boldsymbol{\theta}_M. \quad (2.14)$$

The CRPS on the other hand is defined in terms of predictive CDF given by

$$\text{CRPS}_{x,M+h} = \int \left[F(z|Z_{1:M}) - \mathbb{1}_{Z_{x,M+h} \leq z}\right]^2 dz, \tag{2.15}$$

where $\mathbb{1}$ denotes the indicator function and $F(z|Z_{1:M})$ the CDF of the predictive density $p(z|Z_{1:M})$. Evaluation of (2.14) and (2.15) requires replacing the PDF and CDF of the predictive density with their empirical counterparts obtained using samples from the posterior. Both scoring rules are available in the R package `scoringRules` (Jordan et al., 2019). When deciding between two competing models, the one with the lower total score summed over all future observations and age groups is considered the better choice.

In addition to scoring rules, we also compare the mean squared error (MSE) and mean absolute error (MAE) of our posterior mean forecasts of the log mortality rates using Equation (2.13) and the observed log mortality rates.

2.6 Data analysis during COVID-19: In-sample performance

Our primary objective is to introduce an enhanced Lee-Carter model capable of accurately capturing the fluctuations in log death rates driven by the COVID-19 pandemic. The classical approach of assessing predictive accuracy through data splitting (training and testing data via out-of-sample validation) faces unique challenges in our context. The pandemic predominantly impacts the most recent years of data, making it impractical to exclude these years and estimate parameters using only the earlier data. Such an approach would overlook the pandemic's specific dynamics.

Given these challenges, our analysis of the COVID data focuses solely on evaluating the in-sample fit of the models. To demonstrate the effectiveness of our refined Lee-Carter approach, we will showcase its performance in three distinct countries: the United States, Spain, and Poland. These countries were chosen as illustrative examples due to their significant experiences with the COVID-19 pandemic and the availability of relevant mortality data. In the following sections, we will delve into the model's parameter estimates and its capacity to capture the unique mortality patterns observed during the pandemic in each of these nations.

The data used for this study was mainly sourced from the Human Mortality Database (HMD)[2] and Eurostat[3]. We focused on three western countries that experienced significant COVID-19 impacts, as indicated by deaths per 100,000 population[4]. Specifically, we analysed unisex populations from the United States (US), Spain and Poland.

[2]HMD website: https://www.mortality.org/
[3]Eurostat website: https://ec.europa.eu/eurostat/de/web/main/data/database
[4]COVID-19 data: https://coronavirus.jhu.edu/data/mortality

For Spain and Poland, we obtained data from Eurostat, however, for Poland, data was available starting in 1990. To create a consistent dataset we selected this to be the starting year for all countries in question. We obtained yearly counts of death from Eurostat up until 2022, and combined them with provisional counts of weekly deaths for 2023, aggregating the latter to obtain annual death counts. Exposure was available from Eurostat until 2023 for Spain, however, only until 2020 for Poland. Population estimates of the latter for 2021 - 2023 were obtained from Statistics Poland [5]. In the case of the US, we acquired death and exposure counts up to 2021 from HMD. For the years 2022 and 2023 we obtained provisional counts of deaths and exposure from the National Center for Health Statistics (NCHS) and provisional estimates of the mid year population from the United States Census Bureau. Both were available to download from the Centers for Disease Control and Prevention (CDC) website[6]. To be able to combine multiple data sets by country, we had to go with the lowest granularity regarding age group size. Provisional counts of deaths for the US were given for age groups of width 10, except for the youngest age group. Hence, for sake of comparison, we choose to adopt this as the general age structure for all countries. Meaning that the data was organised into a total of $A = 10$ age groups from '< 5', '5-14' up until '85+'.

A summary of the data sources can be found in Table 2.1.

Table 2.1: Sources of mortality data

Country	Year	Source
Counts of Death		
Poland	1990 - 2023	Eurostat
Spain	1990 - 2023	Eurostat
United States	1990 - 2021	HMD
	2022 - 2023	CDC
Population Estimate		
Poland	1990 - 2020	Eurostat
	2021 - 2023	Statisics Poland
Spain	1990 - 2023	Eurostat
United States	1990 - 2021	HMD
	2022 - 2023	US Census Bureau

For the COVID data, we assume $N_T = 0$ for the US, Spain and Poland in order to ensure identifiability of the parameters for our AR and MA model.

[5]Statistics Poland: https://stat.gov.pl/en/topics/population/population/
[6]CDC data: https://wonder.cdc.gov/mcd-icd10-provisional.html

2.6.1 United States

We applied our models to the US mortality data spanning from 1991 to 2023, comparing it with the Liu-Li model. When evaluating the goodness of fit using the WAIC and LOO-CV metrics, our proposed models demonstrate superior performance (see Table 2.2). More precisely, the MA offers the best in-sample fit, according to our metrics, followed by the AR model. However, the difference between all three models is moderate, which is not unexpected given that the models diverge mainly over a period of three years after the start of the COVID pandemic, i.e. 2021-2023.

Table 2.2: In-sample fit comparison of the Liu-Li and own models on COVID-19 data for multiple countries. Bold value denotes best of the column.

Model	US	Spain	Poland
WAIC			
AR	-1463.47	-1193.73	-1174.92
MA	**-1465.79**	**-1194.63**	**-1181.99**
Liu-Li	-1462.45	-1187.28	-1178.66
LOO-CV			
AR	-1455.42	-1188.21	-1166.10
MA	**-1458.00**	**-1188.66**	**-1176.04**
Liu-Li	-1453.39	-1180.85	-1170.29

Looking at the parameter estimates of the MA model in more detail, we notice that the model demonstrates remarkable confidence in its assessments of the data for 2020 and 2021, with posterior mean estimates of N_{2020} and N_{2021} equalling 1 and a posterior standard deviation of 0. This high level of confidence indicates that the model considered these years as shock years with absolute certainty. Conversely, the other years showed extremely low to negligible posterior means, making them unsuitable candidates for jump years in the model. It should be noted, that this denotes one of the advantages of using a Bayesian approach to parameter estimation, in that N_t is treated as a parameter rather than a random variable, allowing us to explicitly analyse the estimated timing of a mortality shock. For an overview of the estimated occurrences of jump years, see panel a) of Figure 2.1. Another set of interesting parameters denote those of the jump effect J_t, namely μ_Y and σ_Y and b (see Figures 2.2 and 2.3). In the MA model, these parameters yield posterior mean estimates, with values of 1.15 (80%-PI $[0.24, 2.02]$) and 1.27 (80%-PI $[0.37, 2.47]$), respectively. The moving average parameter b is estimated to be medium-sized with a posterior mean of 0.5 (80%-PI $[0.33, 0.68]$). It's worth acknowledging the relatively large standard deviation in the estimates. This variability is not unexpected, as the model predominantly considers only two years extreme events, meaning that parameter estimation is based on these two years only. Furthermore, in such data-scarce

scenarios, the prior distribution significantly influences posterior estimates, especially via the choice of hyperparameters.

As an example for a single country, we can compare the posterior estimates of b with that of the AR parameter, a. Here we see a slightly lower posterior mean for a of 0.39 (80%-PI $[0.30, 0.48]$). Intuitively, this makes sense, since in the AR model the jump parameter J_t is given as a linear combination of all of the past jump effects and not just the immediate past. However, regardless of the model choice, the estimated parameters a and b suggest a substantial amount of serial dependence in the data that is not captured by the Liu-Li approach, as indicated by the probability of b and a being greater than 0.1 being 99.6% and 99.7% respectively, and their maximum a posteriori (MAP) values being 0.48 and 0.41 respectively. All posterior estimates of the AR and MA models can be found in Table B.1 and Table B.4 in the Appendix.

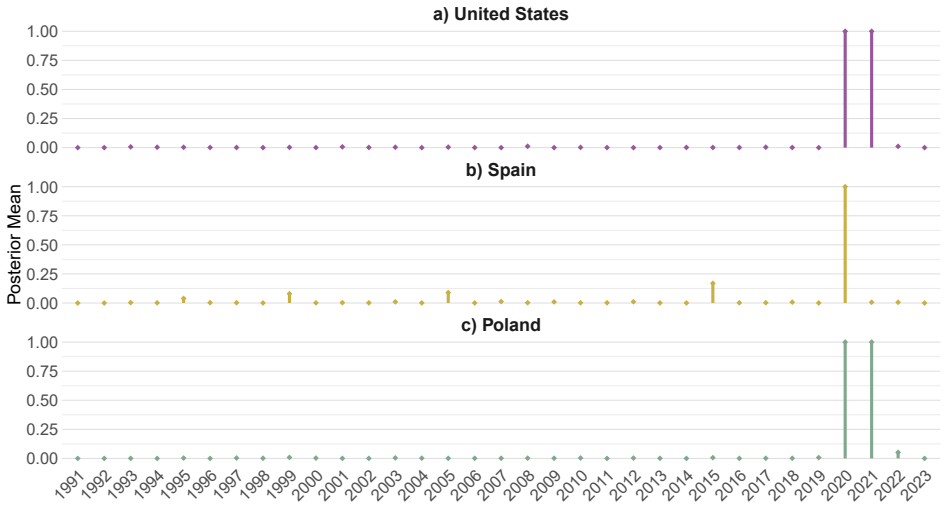


Figure 2.1: Comparison of posterior mean estimates for N_t across time for all countries.

The posterior mean estimates of μ_Y and σ_Y underscore the substantial impact of the COVID-19 pandemic on mortality rates in the US. However, this influence appears to be more diffuse, affecting a broader age spectrum rather than concentrating on specific age groups. Intriguingly, the mortality jump pattern, as represented by $\beta_x^{(J)}$, exhibits a plateau in the middle age range (from 15-24 to 45-54), with a milder impact observed at higher ages. This pattern aligns with findings from Faust et al. (2022), which revealed the most significant relative increase in mortality among the 18-49 age group, corresponding to the working population that played a central role in the spread of COVID-19 (Monod et al., 2021). The posterior mean estimates along with their 80%-PI of $\beta_x^{(J)}$ are depicted in Figure 2.4 across all age groups.

2.6.2 Spain

Comparing the model fit of Liu-Li and our models using the WAIC and LOO-CV, we again see a superior performance of the models that allow for a serial dependent jump component (cf. Table 2.2). The WAIC of both the AR and the MA model is very comparable, however the MA model has the lower estimate.

Both the Liu-Li model, as well as the two other models estimate the posterior probability of a jump occurring in 2020, i.e. N_{2020} to be one, while most of the other years, including 2021, have a negligible posterior mean estimate close to zero. However, the effect of the pandemic does not disappear in 2021 as the vanishing effect b has a posterior mean of 0.21 (80%-PI $[0.1, 0.32]$), indicating a medium sized effect (cf Figure 2.3). This means, that on average 22% of the shock in 2020 is still present in 2021.

The severity of the COVID effect in Spain is on a similar level to that of the US, with a posterior mean estimate of μ_Y given by 1.19 (80%-PI$[0.17, 2.52]$), while that of σ_Y is 1.34 (80%-PI$[0.33, 2.84]$). Posterior distributions can be found in panel b) of Figure 2.2. Looking at the age pattern of the posterior shock, it is clear that the pandemic had a greater impact on older age groups than on younger ones. Two notable things are seen for the posterior estimates of $\beta_x^{(J)}$ depicted in Figure 2.4. First, there is a small plateau in the posterior mean for middle ages. At higher ages, the posterior mean then decreases, only to increase again at the end of the age spectrum, reaching its global maximum in the 75-84 age group. Thus, both the mortality rates of the medium aged as well as the elderly were affected most by the COVID-19 pandemic. Posterior estimates of all parameters, including uncertainty quantification for both the AR and MA model can be found in Table B.2 and B.5 in the Appendix.

2.6.3 Poland

Comparing the model fit of all three models using the WAIC and LOO-CV, we again see superior performance of a model that captures the serial dependence of the COVID-19 pandemic (c.f. Table 2.2). However, for Poland, only the MA model does better than the Liu-Li approach in terms of in-sample fit. This can be explained by the following observation in that the COVID-19 pandemic still has a substantial effect in 2022 and disappears completely in 2023. The specific structure of the MA model allows such a pattern to be captured, while the AR model still accounts for the COVID shock, which affects mortality rates in 2023. The difference in WAIC is thus explained by the difference in model fit for 2023. Looking at the WAIC estimates for 2023 only, we see very comparable values for the Liu-Li and MA models, while the AR model's value is significantly lower.

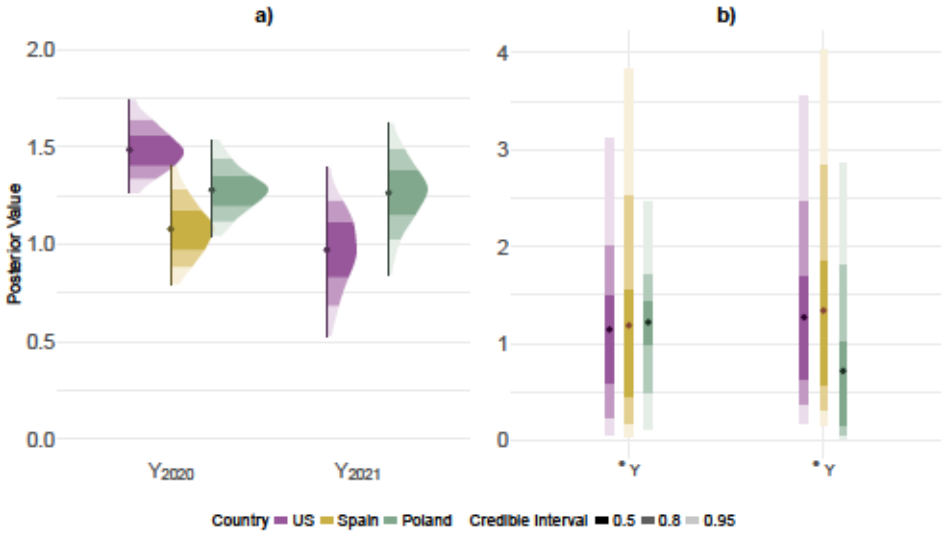


Figure 2.2: Comparison of posterior distribution for the jump parameters Y_t (in a) as well as μ_Y and σ_Y (in b) across countries. The point plotted below or within the density represents the posterior mean.

Similar to the United States, the MA model assumes the years 2020 and 2021 to be jump years, as evident from the posterior estimates of N_{2020} and N_{2021} with mean values of one (cf. Figure 2.1). However, the effect of the pandemic was still present in 2022, as indicated by the estimate of the MA parameter b with a posterior mean of 0.5 (80%-PI[0.363, 0.655]) shown in Figure 2.3. Moreover, the size of the COVID shock was similar in severity to that of the other countries, with a posterior mean for μ_Y of 1.19 (80%-PI[0.3, 1.94]), while σ_Y is estimated to be 1.16 (80%-PI[0.128, 2.6]). The posterior distributions can be found in panel b) of Figure 2.2.

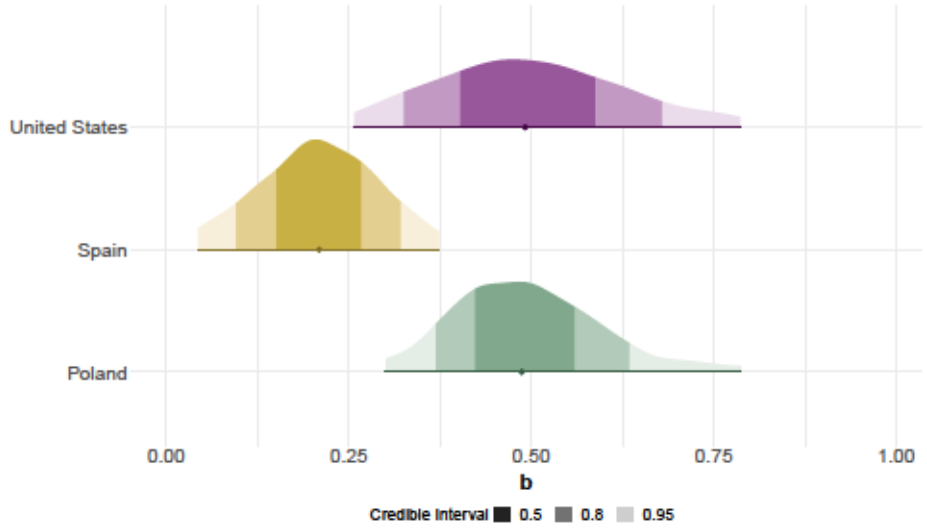


Figure 2.3: Comparing posterior distributions of b across countries. The point below the density represents the posterior mean.

Especially, the older population has been impacted the most by the COVID-19 pandemic

as evident from the posterior estimates of $\beta_x^{(J)}$ depicted in Figure 2.4. Here, we see the lowest posterior mean for the young children aged 5-14, as well as a linear increase for higher ages with a maximum for the age group 75-84. Posterior estimates of all parameters, including uncertainty quantification, for both the AR and MA model can be found in Table B.3 and B.6 in the Appendix.

2.6.4 Comparisons of pandemic effects across countries

After having obtained estimates, we can compare the results across countries. First, and most notably, the MA model consistently performs best, as evident by the lowest values of WAIC and of WAIC and LOO-CV (cf. Table 2.2). We should note, that we also fit a MA-(2) model to all countries, however without improvement in the in-sample fit compared with the MA-(1) model. We therefore refrain from showing the results. The AR model, a choice where the initial shock affects subsequent periods longer, performs better than the Liu-Li model only for two the US and Spain. However, it can still be considered a good alternative. In terms of parameter estimates, the ones of the AR and MAR model are fairly similar. However, for sake of brevity we have decided not to show these results in detail.

Comparing posterior parameter estimates across various countries provides insights into the extent of their exposure to the COVID-19 pandemic. A compelling example is illustrated in Figure 2.1, depicting the jump occurrences of all three countries. Notably, for Poland and the US, the initial pandemic shock has lasted for multiple periods, i.e., 2020 and 2021, before slowly vanishing, while for Spain it denotes a one-period event. The results are in line with those of other researchers who investigated excess mortality rates in Europe during the COVID pandemic and found high values in Spain for the year 2020 and Poland for both 2020 and 2021 (Bonnet et al., 2024). Looking at the impact of COVID, we can compare the estimates of the intensity Y_{2020} across all countries and see that the US has been affected the most followed by Poland and Spain. However, the impact of the second COVID wave in 2021 has been the highest in Poland. The posterior estimates of Y_t can be found in panel a) of Figure 2.2.

Next to magnitude of the COVID pandemic, we can also compare the hyperparameters of the jump intensity by looking at the posterior estimates μ_Y and σ_Y for each country (see panel b) of Figure 2.2). The posterior mean of μ_Y is comparable across all countries, however the underlying uncertainty surrounding these estimates is distinct. For example, in Spain, we see that the posterior distribution of μ_Y is the widest, which is due to there being a single jump only. With more jumps, the posterior gets more weight by likelihood thereby decreasing uncertainty, as evident in Poland. Here, the posterior distribution is the sharpest because there are multiple jumps with similar intensity. The US is somewhere

between these extremes. For σ_Y, we observe comparable results in that the estimate for Spain has the most variation and that of Poland the least.

Furthermore, leveraging the estimated vanishing effect parameter b, we can assess the pace of recovery from shocks across countries. As demonstrated in Figure 2.3, a distinct recovery pattern emerges. In the case of Spain, the posterior estimate is moderate, whereas those for Poland and the United States can be considered high. The data indicates that Spain has undergone a more rapid recovery, while the United States and Poland continue to face increased mortality due to COVID-19 even in later years.

Lastly, the shock pattern exhibited by $\beta_x^{(J)}$ reveals distinct variations across different countries. To illustrate, in Poland, the shock predominantly affects the older population, whereas in the United States, the impact is distributed across a wide range of age groups, displaying a pronounced peak within middle age groups, as indicated by the posterior estimates. Spain occupies an intermediate position between the aforementioned patterns. This variance is visually depicted in Figure 2.4, where the posterior mean shock pattern, along with its 80%-PI, is shown for each country across all age groups.

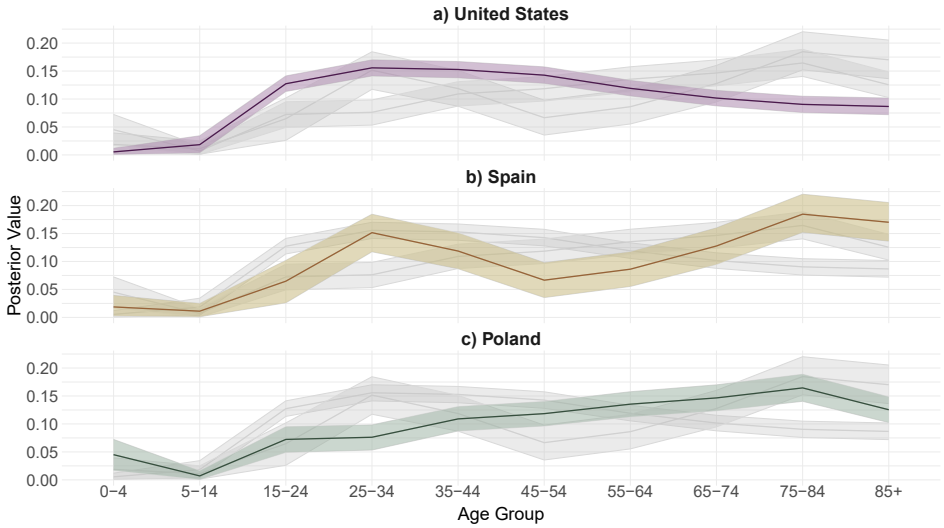


Figure 2.4: Posterior estimates of the jump effect $\beta_x^{(J)}$ for each country. Thick line denotes posterior mean while the shades denote 80%- posterior intervals. Gray shaded area shows respective estimates of the other countries.

2.6.5 Measuring the shock effect by age groups

After having estimated all the parameters, we were able to forecast future death rates. Moreover, we can calculate how much future death rates are affected by the addition of the age-specific jump effect $\beta_x^{(J)} J_t$. Using draws from the posterior predictive distribution, we can calculate empirical quantiles for the shock component $\beta_x^{(J)} J_t$ to answer the question of how much of a percentage increase in death rates is likely to occur in the future due to a shock.

Hereby, consider that the log death rate of our model consists of the log death rate of a LC model, i.e. a scenario without a jump, denoted $\ln\left(m_{x,t}^{LC}\right)$, plus the shock component $\beta_x^{(J)} J_t$ in case there is a jump:

$$\ln\left(m_{x,t}\right) = \underbrace{\alpha_x + \beta_x \kappa_t + e_{x,t}}_{\ln\left(m_{x,t}^{LC}\right)} + \underbrace{\beta_x^{(J)} J_t}_{\ln(c_{x.t})}. \tag{2.16}$$

Using basic rules of logarithms, we can calculate the percentage increase induced by the shock component to the death rate of a jump free scenario, with

$$m_{x,t} = m_{x,t}^{LC} \cdot c_{x,t}. \tag{2.17}$$

In Figure 2.5, we have plotted 90%, 95%, and the 99% credible interval of the predicted percentage increase $\tilde{c}_{x,t} = \exp(\beta_x J_t) - 1$ by age group averaged over time. In addition we have added a dashed line showing the actual, observed increase in death rates due to the COVID-19 pandemic. Here, we have taken the average age specific death rates from 2016 to 2019 and calculated by how much this average was increased (or decreased) in 2020. We note that this increase is given in percentages, not percentage points.

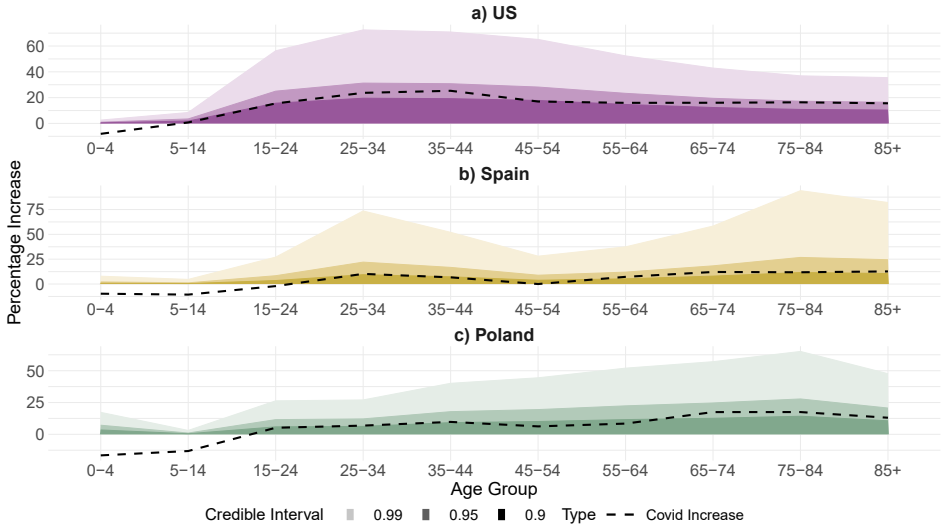


Figure 2.5: Observed percentage increase in death rates from the average of 2016 - 2019 to 2020 by countries (COVID increase). The different shades denote the respective width of the prediction interval for $\bar{\tilde{c}}_x = \frac{1}{T}\sum_{t=1}^{T}\left[\exp\{\beta_x J_t\} - 1\right]$.

First, looking at Figure 2.5 we can see that the COVID-19 pandemic did not induce a parallel constant shift in the log death rates, as some authors (e.g. Schnürch et al., 2022) proposed. Adding a constant to all age-specific log death rates would result in a constant percentage increase of the mortality rates, which is not observed. Second, using Figure 2.5, we can provide an upper bound on the increase in mortality rates that our model predicts for a future time period. For example, our model states that for any future single

year, the mortality rates for the age group 75-84 in Spain will not increase by more than 10% with a probability of 95%. However, we are aware that it is difficult to draw general conclusions for a future pandemic after having observed just one. The results should not be considered an attempt to forecast the severity of a future pandemic but rather as a tool to better visualise and capture the impact of the mortality shocks in the past.

2.7 Data analysis during the world wars: out-of-sample performance

As mentioned in Section 2.6, the recent occurrence of the pandemic does not allow for comparison of forecasting accuracy on out-of-sample data. Moreover, due to the rare nature of a mortality shock, i.e. there has not been another pandemic of this magnitude in the data set, we have to set the last jump indicator N_T to zero to ensure model identifiability (see section B.1 in the Appendix), which makes forecasting during the pandemic difficult. There is however mortality data available for multiple countries in Europe during both World War I and World War II. Both wars, as well as the Spanish flu in 1918, can be considered as mortality shocks. We obtained mortality and population data for England and Wales from 1900 - 2000 from the HMD website and, for sake of consistency, divided the data into the same ten age groups as for the COVID data. We note that this is the same data used by Liu and Li (2015).

We compare the out-of-sample performance of our models on the war data using a training period (TRP) and a test period, distinguishing between two scenarios. Scenario one trains the parameters on the First World War and then predicts the future behaviour of mortality rates during the Second World War. Scenario two predicts future death rates after both world wars have occurred, that is, during a time of low volatility. In the second scenario, we aim to determine if specifically accounting for the two wars, that is, the mortality shocks, can reduce the uncertainty of the non-jump parameters and, thus, the uncertainty of our forecasts, without treating the extreme observations as outliers and removing them from the data, as is done, for example, by Lee and Carter (1992). This is of particular interest for forecasting mortality in the years following the COVID pandemic.

2.7.1 Prediction of future death rates during times of war

Following the methodology of Section 2.5.2, we can forecast future mortality rates recursively by applying Equation (2.13). We have trained the AR, MA, and Liu-Li models on the England and Wales data for the years 1901 - 1943 and predict future death rates for a total of $H = 5$ years ahead. In this scenario, we specifically choose to compare short-term forecasts, as the models' predictions will only differ significantly in the near future, and

we do not want to dilute the results. For longer-term forecasts, all models will produce similar predictions, as shown in the second scenario below. Moreover, in this scenario, we do not have to assume that $N_T = 0$, since the vanishing parameters can be estimated during the period of the First World War, and we know of a time $\tilde{t}$ where there is no jump, e.g., 1930 (cf. section B.1).

Examining the predictive performance of the Liu-Li model against the models that include some form of serial dependence in the jump component, we find that the latter models outperform the former by virtue of a lower score. In addition, the AR model, which carries the initial shock the longest in the future, performs best, an unsurprising result given the timing of the last training observation. Out-of-sample results can be found in Table 2.3.

Table 2.3: Out-of-sample fit comparison of the Liu-Li and our models on England and Wales data. Bold value denotes the best model for each column.

Score	AR	Liu-Li	MA	LC
TRP = 1901-1943, H = 5				
LogS	**-26.39**	-24.00	-24.92	-
CRPS	**6.36**	6.84	6.69	-
MSE (in percentages)	**8.55**	10.51	9.44	-
MAE (in percentages)	16.64	17.47	**16.60**	-
TRP = 1901-1980, H = 30				
LogS	**-68.00**	-67.48	-67.78	128.46
CRPS	32.59	32.74	**32.55**	68.25
MSE (in percentages)	**3.67**	3.73	3.71	15.61
MAE (in percentages)	**14.49**	14.63	14.61	31.72

Notes: TRP = training period. H = forecast horizon.

2.7.2 Prediction of future death rates during normal times

In the second scenario, we trained all models on the England and Wales data for the years 1901 - 1980. In addition, we fit a standard LC model as well, where the parameters were estimated on the mortality improvement rates using the Route II approach. For all models, we created long term forecasts, i.e., future death rates for $H = 30$ years ahead. As shown in Table 2.3, the models that explicitly account for the past mortality shocks, namely, the Liu-Li, AR, and MA models, produce superior forecasts in terms of lower scores. More precisely, the predictive accuracy of the jump models during times of low volatility is almost identical and the slight deviation can be attributed to sampling variation, as the results are all within less than 1% of each other. Thus, in terms of forecasting, there is nothing to lose by choosing a more complex model, such as the AR model, over the Liu-Li model, even when forecasting in normal times. On the other

hand, the predictive accuracy of the LC model is considerably lower, as the mortality shocks affect the parameter estimates of β_x and d substantially, while also increasing the variance of $(\xi_t)_t$ and $(\varepsilon_{x,t})_t$. Both effects significantly reduce the predictive power and highlight the need to explicitly account for past unusual mortality events, such as wars or pandemics, if accurate estimates of future mortality rates are to be obtained.

To visualise the above results, we compare the forecasts of the LC and our model in Figure 2.6. Here, we notice the substantially larger prediction intervals of the LC model compared with those of the AR model.

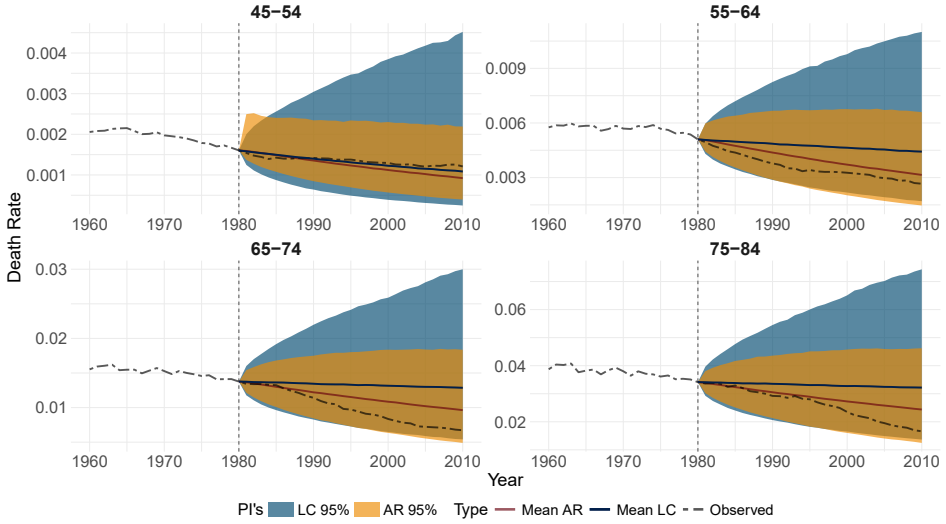


Figure 2.6: 30-year ahead forecasts of death rates for different age groups of England and Wales including prediction intervals (PI) using both a LC and AR model. Dashed black line denotes actual observed values, while blue line denotes the mean forecast (Mean LC) of the LC model and red line the mean forecast (Mean AR) of own model.

Finally, we point out that prediction intervals of future death rates are used in an insurance context to determine the solvency capital for life insurance liabilities (see e.g. Robben and Antonio (2024)). Figure 2.6 illustrates that the prediction intervals from the Lee-Carter model are unrealistic and there is an important need to build a mortality model that adequately accounts for mortality shocks such as the models proposed in this paper.

2.8 Extension: Multi-population mortality model with vanishing jumps

Instead of estimating the model for single populations only, so-called multi-population models provide more robust mortality forecasts for multiple populations that share similar socioeconomic characteristics. Their strength lies in the usage of more data by combining different data sources, i.e. countries, which increases stability. Li and Lee (2005) were

the first to extend the classical LC model into a multi-population framework. A Bayesian implementation was introduced by Antonio et al. (2015). For an overview, we refer to Enchev et al. (2017).

There are many possibilities for extending our proposed modelling framework into a multi-population setting. A detailed discussion of all of them would exceed the scope of this paper. Examples include a common factor model, where all countries share a common age and time effect (e.g. Li & Lee, 2005). However, the addition of another age-time interaction term increases the amount of identifiability constraints needed. Alternatively, a two-step estimation approach may be considered (Antonio et al., 2015). Another possible choice is the co-integrated LC model (e.g. Li & Hardy, 2011), where the time parameter κ_t follows a multivariate random walk with drift.

In the following, we propose a possible multi-population extension with a shared mortality jump occurrence parameter given by the following model

$$\ln(m_{x,t,c}) = \alpha_{x,c} + \beta_{x,c}\kappa_{t,c} + \beta^{(J)}_{x,c} J_{t,c} + \varepsilon_{x,t,c} \tag{2.18}$$

with $c = \{1, 2, \ldots, C\}$ denoting the index for a given country, while $\boldsymbol{\kappa}_t = (\kappa_{t,1}, \ldots, \kappa_{t,C})^\top$ is modelled as a multivariate random walk with drift. We can use both the AR as well as the MA structure to model the serial dependence of the country specific jump component $J_{t,c}$. The model of (2.18) using an AR structure is given by

$$J_{t,c} = a_c \, J_{t-1,c} + N_t \, Y_{t,c}.$$

Here, N_t implies the existence of a global shock that affects all countries with a local effect size $J_{t,c}$. Similar to single-population models proposed in Section 2.2, each country has its own vanishing parameter a_c and jump severity $Y_{t,c}$. However, they share the same jump occurrence process N_t.

Note that all identification conditions derived for single population models directly carry over to the present setting. The only thing that needs to be checked is whether a model of the form (2.18), where $(\beta_{x,c})_x$ or $(\beta^{(J)}_{x,c})_x$ do not sum to 1, can be transformed into a model that satisfies these conditions without deviating from the joint value of the global parameter N_t. This is easily achieved by appropriate rescaling of the country-specific parameters, that is by using $\beta_{x,c}/(\sum_x \beta_{x,c})$, $\beta^{(J)}_{x,c}/(\sum_x \beta^{(J)}_{x,c})$, $\kappa_{t,c} \sum_x \beta_{x,c}$, and $Y_{t,c} \sum_x \beta^{(J)}_{x,c}$ instead of $\beta_{x,c}$, $\beta^{(J)}_{x,c}$, $\kappa_{t,c}$, and $Y_{t,c}$, respectively.

2.8.1 In-sample comparison

To compare single and multi-population models, we use the WAIC metric. When fitting single-population models separately, it is implicitly assumed that the parameters and likelihoods of the single-population models are independent of each other. Therefore, we want to compare the WAIC of the joint multi-population model, which accounts for the dependence between countries, with the sum of the WAIC scores of the single-population models. Indeed, when assuming independence, the total WAIC can be decomposed into the sum of the single-population WAICs, as shown below.

Suppose some data y with total sample size n can be decomposed in two parts, where $I_1 = \{1, \dots, m\}$ and $I_2 = \{m+1, \dots, n\}$, with $i \in I_1$ and $j \in I_2$ as well as corresponding parameters $\theta = (\theta_1, \theta_2)^\top$ that are independent of each other. Moreover let θ_1 denote the parameters of all y_i and θ_2 those of all y_j. The lpd can then be approximated as

$$\begin{aligned}
\widehat{\text{lpd}} &= \sum_{l=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\big(y_l|\theta^{(s)}\big)\right) \\
&= \sum_{i\in I_1} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\big(y_i|\theta_1^{(s)}\big)\right) + \sum_{j\in I_2} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\big(y_j|\theta_2^{(s)}\big)\right) \\
&= \widehat{\text{lpd}}_1 + \widehat{\text{lpd}}_2.
\end{aligned}$$

In addition, the penalty term of the WAIC for the amount of parameters can be expressed as

$$\begin{aligned}
\widehat{p}_{waic} &= \sum_{l=1}^{n} V_{s=1}^{S} \log p\big(y_l|\theta^{(s)}\big) \\
&= \sum_{i\in I_1} V_{s=1}^{S} \log p\big(y_i|\theta_1^{(s)}\big) + \sum_{j\in I_2} V_{s=1}^{S} \log p\big(y_j|\theta_2^{(s)}\big) \\
&= \widehat{p}_{waic}^{(1)} + \widehat{p}_{waic}^{(2)},
\end{aligned}$$

The above decomposition of both $\widehat{\text{lpd}}$ and $\widehat{p}_{waic}$ allows us to compare the WAIC of the joint multi-population model with the sum of WAIC's of the single-population model.

The LOO-CV can be decomposed similarly granting us another measure to compare the in-sample fit of the single and multi-population models. Moreover, let $y_{I_1} = (y_1, \dots, y_m)^\top$

and $y_{I_2} = (y_{m+1}, \ldots, y_n)^\top$ then we obtain:

$$\begin{aligned}
\text{lpd}_{\text{loo}} &= \sum_{l=1}^{N} \log p(y_l|y_{-l}) \\
&= \sum_{l=1}^{N} \log \int p(y_l|\theta) p(\theta|y_{-l}) d\theta \\
&= \sum_{l=1}^{N} \log \int \int p(y_l|\theta_1, \theta_2) p(\theta_1, \theta_2|y_{-l}) d\theta_1 d\theta_2 \\
&= \sum_{i \in I_1} \log \int p(y_i|\theta_1) p(\theta_1|y_{-i}) d\theta_1 \underbrace{\int p(\theta_2|y_{I_2}) d\theta_2}_{=1} + \\
&\quad \sum_{j \in I_2} \log \int p(y_j|\theta_2) p(\theta_2|y_{-j}) d\theta_2 \underbrace{\int p(\theta_1|y_{I_1}) d\theta_1}_{=1} \\
&= \text{lpd}_{\text{loo}}^{(1)} + \text{lpd}_{\text{loo}}^{(2)}.
\end{aligned}$$

2.8.2 Results

We applied the multi-population model to the original mortality data from 1991 to 2023 to estimate the parameters for the United States, Spain, and Poland jointly. For the multi-population model, the posterior mean estimates of the jump occurrences N_{2020} and N_{2021} are both estimated to be 1. That is, the model assumes that both of these years are shock years, similar to the single-country estimates for the US and Poland. Looking at the WAIC for the AR model, the in-sample fit is drastically improved by the multi-population approach (see Table 2.4). The lpd of both the single- and multi-population models is very comparable, however the latter uses $3 \cdot (T-3) - 3$ fewer parameters (the N_t parameters, excluding the constraints, for all three countries, minus three correlation parameters for $\Delta\boldsymbol{\kappa}_t$) resulting in a substantially lower WAIC. For the MA model, we observe the same findings: the log scores of the single- and multi-population models are very comparable, but the latter approach uses substantially fewer parameters, resulting in a lower WAIC value. Looking at the LOO-CV metric, we also see superior performance of the multi-population approach, with the MA extension obtaining lower scores than the AR approach.

Another advantage of the multi-population approach lies in its property of a joint jump occurrence N_t. If we were to generate a single forecast using the multi-population approach, a future pandemic occurs at the same time across all countries which is not necessarily the case for independent single-population models.

Table 2.4: In-sample fit comparison on COVID-19 data for the multi-population approach. Bold value denotes best of the column.

Model	Single Population	Multi Population
WAIC		
AR	-3832.12	**-3850.42**
MA	-3841.02	**-3859.42**
LOO-CV		
AR	-3809.73	**-3830.30**
MA	-3819.76	**-3839.80**

2.9 Conclusion

In this paper, we have introduced a new class of models that allows for more accurate modelling of mortality rates in the event of a shock. More precisely, we have extended the well-known LC model structure to include a serially dependent jump effect, where the impact of a shock is largest at the beginning and then gradually diminishes over time, offering considerable flexibility. Compared with the approach of Liu and Li (2015), our models better capture the underlying pattern of the COVID-19 pandemic for Spain, the US and Poland. Additionally, we have demonstrated that the jump auto-correlation structure is applicable to various shock scenarios, as evidenced by the improved out-of-sample fit in the case of war-related data for England and Wales. We have also shown that explicitly accounting for past shocks is crucial for making accurate predictions of future mortality. Moreover, we introduced a multi-population extension with a shared jump occurrence parameter. Finally, we have proven the identification of parameters for both the single-population and multi-population models.

A valid point of criticism, nonetheless, relates to the considerable variability observed in the jump parameters, namely μ_Y, σ_Y, and a, respectively b. Employing a Bayesian hierarchical modelling approach could potentially reduce this variability by pooling information across dimensions. However, the general problem remains: No matter the approach, the model will have difficulty in estimating the parameters with only a few data points available. However, with the availability of an increased number of time points and thus mortality shocks, a substantial reduction in the standard deviation will be achieved. In general, our method's efficacy is most pronounced when a larger temporal scope is available, enabling more robust estimation of the parameters.

On the other hand, an advantage of the Bayesian approach that we have not explored further is the use of expert opinion for the specific choice of hyperparameters in prior distributions. For example, Zhou and Li (2022) incorporates expert opinion to simulate future mortality scenarios for events similar to the COVID outbreak. The same expert

opinion could be used to set more informative priors, exploiting the interplay between prior and likelihood to update the posterior. In data-rich scenarios the estimates are primarily influenced by the data, whereas in data-scarce settings, the influence of the prior, or expert opinion, becomes more pronounced. This approach may substantially reduce the posterior variability and can be seen as a middle ground between purely expert based and data-only estimation.

Moreover, our models assume that after a shock, the trend of mortality rates tends to return to their pre-shock trend driven by their constant drift. However, it is also possible that the shock has introduced either a new trend or baseline level. For example, after a severe pandemic, the population may be more alert to infectious diseases, leading to greater caution during the winter months. This change in behaviour could lead to lower levels of mortality after the pandemic. On the contrary, the impact might decrease but not disappear completely. Rather, it could converge toward a general baseline level, which results in a permanent effect that can be compared to other causes of death such as the flu. Such scenarios which are briefly discussed in van Berkum et al. (2025) are not considered by our approach but can constitute an interesting generalisation of our model.

Furthermore, the shock of a pandemic's mortality can trigger a compensatory response. There is an argument that a pandemic accelerates the demise of those already in poor health. This type of scenario, while not observed, was especially discussed at the beginning of the COVID-19 pandemic (e.g. Cairns et al., 2020). Here it is believed that many of those who die during a pandemic would have died anyway in the near future, resulting in a slight decrease in the mortality rates among survivors. This contradicts our assumption of a pandemic effect that slowly vanishes over time, making our model unsuitable for this type of scenario. In practical applications, our model holds promise for actuarial contexts, particularly in determining solvency capital for mortality and longevity risk, which is imposed by supervisory authorities. The wider confidence intervals provided by our approach suggest that insurance companies may need to increase their capital reserves to safeguard against future pandemics and mortality shocks. This highlights the real-world significance and potential impact of our modelling framework on risk management in the insurance industry.

Acknowledgment

Julius Goes would like to thank the Bamberg Graduate School of Social Sciences (BAGSS) for their support. Moreover, the authors thank the two reviewers and the associate editor for helpful and constructive comments that led to a much improved manuscript.

Funding

Julius Goes gratefully acknowledges financial support by the Oberfrankenstiftung (grant FP01054).

Data availability

For full replication of the results, we provide the code including data at our GitHub repository, available at https://github.com/goesj/VanishingJumps

B Appendix

B.1 Proof of Identification

Identifiability of σ_r^2 is obvious. The conditions $\Delta J_2 = 0$, $\xi_2 = 0$ and $\sum_x \beta_x = 1$ yield $\sum_x \mathbb{E}(Z_{x,1}) = d$, i.e. identifiability of the drift, which in turn implies that $\Delta\kappa_2 = d$. Starting from (2.11) with $\mathbf{B_x} = (\beta_x, \beta_x^{(J)})^\top$ and $\mathbf{K_t} = (\Delta\kappa_{t+1}, \Delta J_{t+1})^\top$ we can verify identifiability of $(\beta_x)_x$, $(\beta_x^{(J)})_x$, $(\Delta\kappa_t)_t$ and $(\Delta J_t)_t$ showing that the only possible choice for the matrix $\mathbf{A}$ is the identity matrix if the constraints

$$\sum_{x=1}^{A} \beta_x = \sum_{x=1}^{A} \tilde{\beta}_x = 1, \quad \sum_{x=1}^{A} \beta_x^{(J)} = \sum_{x=1}^{A} \tilde{\beta}_x^{(J)} = 1, \quad \Delta J_1 = \Delta \tilde{J}_1 = 0, \quad \Delta\kappa_1 = \Delta\tilde{\kappa}_1 = d. \tag{B.1}$$

are met. Then, from

$$\begin{pmatrix} d \\ 0 \end{pmatrix} = \tilde{\mathbf{K}}_\mathbf{1} = \begin{pmatrix} a_1 & a_3 \\ a_2 & a_4 \end{pmatrix} \mathbf{K}_1 = \begin{pmatrix} a_1 & a_3 \\ a_2 & a_4 \end{pmatrix} \begin{pmatrix} d \\ 0 \end{pmatrix}$$

we get $a_1 = 1$ and $a_2 = 0$ in view of our assumption $d \neq 0$. Similarly,

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \sum_x \tilde{\mathbf{B}}_x = \frac{1}{a_1\, a_4 - a_2\, a_3} \begin{pmatrix} a_4 & -a_3 \\ -a_2 & a_1 \end{pmatrix} \sum_x \mathbf{B}_x = \frac{1}{a_1\, a_4 - a_2\, a_3} \begin{pmatrix} a_4 & -a_3 \\ -a_2 & a_1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

yields

$$a_4 - a_3 = a_1 - a_2 = a_1\, a_4 - a_2\, a_3.$$

Hence, we end up with $a_3 = 0$ and $a_4 = 1$. In other words, we achieved identifiability of the parameters of the mortality improvements. If we assume additionally that $J_1 = J_2 = 0$, we can identify all of $(J_t)_t$ iteratively using $(\Delta J_t)_t$ as well as $\Delta J_2 = 0$.

We will discuss identifiablity of the parameters of the two proposed time series models for the jump processes separately.

Identification of Parameters in AR(1) Model

To identify all other parameters in the AR(1) model, we have to assume that $N_T = 0$ and that $(Y_t)_t$ is a sequence of positive parameters. This allows us to identify the times of jump occurences $(N_t)_t$, the autoregressive parameter a and jump intensity $(Y_t)_t$. Note that if there is no jump, then the parameter a cannot be identified and can be neglected in the model. Otherwise we can proceed iteratively: Let t^* denote the time of the first jump event, then we have $N_t = 0, J_t = 0$ for all $t < t^*$ as well as $N_{t^*} = 1$ and $Y_{t^*} = J_{t^*}$. Noting that a jump event at time t^* implies that $J_t > 0$ for all $t \geq t^*$ if $a > 0$, we can

deduce that $a = 0$ for $J_t = 0$ for some $t > t^*$. For $a \neq 0$, we have $J_T = aJ_{T-1}$ as $N_T = 0$ by assumption. Hence, we can identify a and subsequently $(N_t)_t$ and the Y_t's belonging to non-vanishing N_t's.

As a final remark, let us mention that instead of assuming $N_T = 0$ we could also assume that we know a time point $\tilde{t} > t^*$, where no jump occurs. Then the whole argument above can be adapted using $\tilde{t}$ instead of T. Moreover, instead of assuming positivity of $(Y_t)_t$ one can alternatively assume that $J_{T-1} \neq 0$.

Identification of Parameters in MA(1) Model

Identification of parameters in the MA(1) model follows a similar structure. Again, we assume that $N_T = 0$ and that $(Y_t)_t$ is a sequence of positive parameters. Let t^* denote the first year of a jump which can be identified by observation of $J_{t^*} > 0$. Then $N_{t^*} = 1$, $Y_{t^*} = J_{t^*}$ and $N_t = J_t = 0$ for all $t < t^*$. Note that $(N_tY_t)_{t \geq t^*}$ forms a non-homogeneous first order difference equation with time-varying coefficients which can be solved recursively:

$$N_{t^*+h}Y_{t^*+h} = (-b)^h J_{t^*} + \sum_{k=1}^{h}(-b)^{h-k} J_{t^*+k-1}, \quad h = 1, \ldots, T - t^*. \tag{B.2}$$

Hence, if we are able to identify b, all remaining parameters are identified. To this end, recall that $N_T = 0$ by assumption which gives

$$0 = (-b)^h J_{t^*} + \sum_{k=1}^{T-t^*}(-b)^{h-k} J_{t^*+k-1}. \tag{B.3}$$

This means, that a unique root of the polynomial (in b) on the right hand side in the interval [0,1) assures identifiability of b.

For illustrative purposes, we can alternatively identify b using the following step-by-step procedure with t^* being the first year of a jump: First, if $J_{t^*+1} < J_{t^*}$, then we can identify two sub-cases. If $J_{t^*+2} = 0$, it follows that $N_{t^*+1} = 0$ and we can estimate b, by noting that $J_{t^*+1} = bN_{t^*}Y_{t^*}$. However, if $J_{t^*+2} \neq 0$, then either $N_{t^*+1} = 1$ or $N_{t^*+2} = 1$. The latter can be checked by seeing if $J_{t^*+4} = 0$ allowing us to estimate b using J_{t^*+3}. In case of the former, identification of the parameters is achieved in a similar way as described next.

Second, if $J_{t^*+1} > J_{t^*}$, then $N_{t^*+1} = 1$, since $b \in [0, 1)$. We then check if $N_{t^*+2} = 0$ by observation of J_{t^*+4}. If this is indeed the case, J_{t^*+1} can be rewritten to obtain $N_{t^*+1}Y_{t^*+1} = J_{t^*+1} - bN_{t^*}Y_{t^*}$, which can be substituted into J_{t^*+2}. Then we solve for b

and obtain the following quadratic expression

$$b_{1/2} = \frac{-J_{t^*+1} \pm \sqrt{(J_{t^*+1})^2 - 4 \cdot (-N_{t^*} Y_{t^*}) \cdot (-J_{t^*+2})}}{2 \cdot (-N_{t^*} Y_{t^*})}.$$

We obtain a unique solution for b by checking which of b_1 or $b_2 \in [0, 1)$ or by observation of a second shock at some later time period onward.

Again, as in the AR-case, instead of assuming $N_T = 0$, we could also assume that we know a time point $\tilde{t} > t^*$, where no jump occurs. One simply has to substitute the upper index $T - t^*$ of the sum in (B.3) by $\tilde{t} - t^*$. In particular, this leads to a reduction of the degree of the polynomial on the right hand side which in turn simplifies the problem of finding its roots.

B.2 Deriving Dirichlet distributions from Gamma distributions

Let X_i be a random variable from the Gamma distribution with $X_i \sim \text{Gamma}(\alpha_i, 1)$, where $i = 1, \ldots, k$. Further, let

$$Y_i = \frac{X_i}{X_1 + X_2 + \cdots + X_k}. \tag{B.4}$$

Then, the joint density of $Y_1, \ldots, Y_{k-1}$ is

$$f(y_1, \ldots, y_{k-1}) = \frac{\alpha_1 + \cdots + \alpha_k}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} y_1^{\alpha - 1} \cdots y_{k-1}^{\alpha_{k-1} - 1} \left(1 - y_1 - \cdots - y_{k-1}\right)^{\alpha_k - 1}, \tag{B.5}$$

where $y_i > 0, i = 1, \ldots, k-1, y_1 + \cdots + y_{k-1} < 1$. The above joint PDF of $Y_1, \ldots, Y_{k-1}$ happens to be the PDF of a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$ for the random vector $(Y_1, \ldots, Y_k)$, with $y_k = 1 - y_1 - \cdots - y_{k-1}$. For proof see e.g. Ng et al. (2011).

Hence, instead of sampling β from a Dirichlet distribution, we can also sample Gamma distributed random variables $b_1, \ldots, b_A$ and apply the transformation of (B.4) by setting

$$\beta_x = \frac{b_x}{b_1 + \cdots + b_A}, \text{ for all } x \in \{1, \ldots, A-1\}$$

and $\beta_A = (1 - \beta_1 - \cdots - \beta_{A-1})$. This allows us to use a wider variety of samplers, like the multivariate slice sampler of Tibbits et al. (2014), improving convergence.

B.3 Tables of parameter estimates

Parameter Estimates of AR models for COVID Data

Table B.1: Posterior estimates for the US data using the AR model

	Mean	MAP	Sd	10%-PI	90%-PI	split-$\hat{R}$	Bulk-ESS	Tail-ESS
β_1	0.11	0.10	0.02	0.08	0.13	1.00	1996	1852
β_2	0.16	0.16	0.02	0.13	0.19	1.00	2035	1736
β_3	0.14	0.14	0.02	0.11	0.17	1.00	1982	1884
β_4	0.19	0.19	0.02	0.16	0.22	1.00	1970	2050
β_5	0.16	0.16	0.02	0.13	0.19	1.00	1595	1977
β_6	0.08	0.08	0.02	0.05	0.11	1.00	1847	1887
β_7	0.07	0.07	0.02	0.05	0.10	1.00	1975	1941
β_8	0.05	0.05	0.02	0.02	0.08	1.00	2075	2006
β_9	0.02	0.01	0.02	0.00	0.04	1.00	2036	2048
β_{10}	0.01	0.00	0.01	0.00	0.03	1.00	2036	1759
$\beta_1^{(J)}$	0.01	0.00	0.00	0.00	0.01	1.00	1926	1792
$\beta_2^{(J)}$	0.02	0.01	0.01	0.00	0.03	1.00	1825	1782
$\beta_3^{(J)}$	0.13	0.13	0.01	0.11	0.14	1.00	1994	1791
$\beta_4^{(J)}$	0.16	0.16	0.01	0.14	0.17	1.00	1832	1911
$\beta_5^{(J)}$	0.15	0.15	0.01	0.14	0.17	1.00	1616	2006
$\beta_6^{(J)}$	0.14	0.14	0.01	0.13	0.16	1.00	1842	1824
$\beta_7^{(J)}$	0.12	0.12	0.01	0.11	0.13	1.00	1896	1967
$\beta_8^{(J)}$	0.10	0.10	0.01	0.09	0.12	1.00	1920	2006
$\beta_9^{(J)}$	0.09	0.09	0.01	0.08	0.10	1.00	1916	1852
$\beta_{10}^{(J)}$	0.09	0.08	0.01	0.07	0.10	1.00	1827	1549
d	-0.09	-0.09	0.04	-0.14	-0.05	1.00	2166	2004
σ_ξ	0.20	0.19	0.03	0.16	0.25	1.00	1565	1857
σ_r	0.02	0.02	0.00	0.02	0.02	1.00	2028	1925
p	0.06	0.05	0.03	0.02	0.11	1.00	1819	1695
a	0.39	0.41	0.07	0.30	0.48	1.00	1359	1119
μ_Y	1.31	1.39	0.68	0.42	2.07	1.00	1632	1266
σ_Y	1.07	0.27	0.88	0.19	2.28	1.00	218	259

Table B.2: Posterior estimates for Spain using the AR model

	Mean	MAP	Sd	10%-PI	90%-PI	split-$\hat{R}$	Bulk-ESS	Tail-ESS
β_1	0.12	0.12	0.02	0.09	0.14	1.01	2001	1964
β_2	0.16	0.16	0.02	0.14	0.19	1.00	1933	1899
β_3	0.18	0.18	0.02	0.16	0.21	1.00	2071	1964
β_4	0.15	0.16	0.02	0.13	0.18	1.00	1818	1634
β_5	0.11	0.11	0.02	0.09	0.13	1.00	1983	1838
β_6	0.04	0.04	0.02	0.02	0.06	1.00	1998	1938
β_7	0.04	0.05	0.02	0.02	0.07	1.00	1901	1649
β_8	0.06	0.06	0.02	0.04	0.08	1.00	2008	1940
β_9	0.07	0.07	0.02	0.05	0.09	1.00	1865	1849
β_{10}	0.06	0.06	0.02	0.04	0.08	1.00	1676	1889
$\beta_1^{(J)}$	0.02	0.00	0.02	0.00	0.04	1.00	1941	1964
$\beta_2^{(J)}$	0.01	0.00	0.01	0.00	0.02	1.00	2077	1940
$\beta_3^{(J)}$	0.06	0.07	0.03	0.03	0.10	1.00	1888	1938
$\beta_4^{(J)}$	0.15	0.15	0.03	0.12	0.19	1.00	1904	2096
$\beta_5^{(J)}$	0.12	0.12	0.02	0.09	0.15	1.00	1992	1858
$\beta_6^{(J)}$	0.07	0.06	0.02	0.04	0.10	1.00	2208	1957
$\beta_7^{(J)}$	0.09	0.09	0.02	0.06	0.12	1.00	1961	1908
$\beta_8^{(J)}$	0.13	0.13	0.03	0.10	0.16	1.00	2087	1966
$\beta_9^{(J)}$	0.18	0.18	0.03	0.15	0.22	1.00	1518	1608
$\beta_{10}^{(J)}$	0.17	0.17	0.03	0.13	0.20	1.00	2107	1887
d	-0.28	-0.28	0.05	-0.34	-0.21	1.00	2027	1893
σ_ξ	0.32	0.31	0.05	0.26	0.39	1.00	1591	1724
σ_r	0.03	0.03	0.00	0.03	0.04	1.00	1979	1941
p	0.05	0.02	0.03	0.01	0.09	1.00	1601	1887
a	0.29	0.27	0.11	0.15	0.43	1.00	1770	1690
μ_Y	1.19	1.15	1.01	0.21	2.52	1.00	802	967
σ_Y	1.32	0.61	1.00	0.26	2.71	1.00	224	223

Table B.3: Posterior estimates for Poland using the AR model

	Mean	MAP	Sd	10%-PI	90%-PI	split-$\hat{R}$	Bulk-ESS	Tail-ESS
β_1	0.19	0.19	0.02	0.17	0.22	1.00	1684	1726
β_2	0.19	0.19	0.02	0.16	0.22	1.00	2000	1933
β_3	0.11	0.11	0.02	0.08	0.13	1.00	1887	1926
β_4	0.09	0.10	0.02	0.07	0.12	1.00	1883	1825
β_5	0.09	0.09	0.02	0.06	0.11	1.00	1910	1699
β_6	0.08	0.08	0.02	0.05	0.10	1.00	1740	1904
β_7	0.08	0.07	0.02	0.05	0.10	1.00	2102	1826
β_8	0.05	0.05	0.02	0.03	0.08	1.00	1794	1851
β_9	0.06	0.06	0.02	0.04	0.09	1.00	2098	2029
β_{10}	0.06	0.06	0.02	0.03	0.08	1.00	1809	1829
$\beta_1^{(J)}$	0.04	0.05	0.02	0.02	0.07	1.00	2038	1740
$\beta_2^{(J)}$	0.01	0.00	0.01	0.00	0.02	1.00	1959	1924
$\beta_3^{(J)}$	0.07	0.07	0.02	0.05	0.09	1.00	1994	1774
$\beta_4^{(J)}$	0.08	0.08	0.02	0.05	0.10	1.00	1885	1958
$\beta_5^{(J)}$	0.11	0.11	0.02	0.09	0.13	1.00	1987	1857
$\beta_6^{(J)}$	0.12	0.12	0.02	0.10	0.14	1.00	1823	1902
$\beta_7^{(J)}$	0.14	0.13	0.02	0.11	0.16	1.00	2045	1751
$\beta_8^{(J)}$	0.15	0.15	0.02	0.12	0.17	1.00	2016	1966
$\beta_9^{(J)}$	0.16	0.16	0.02	0.14	0.19	1.00	2118	1915
$\beta_{10}^{(J)}$	0.13	0.13	0.02	0.10	0.15	1.00	1917	1888
d	-0.26	-0.26	0.05	-0.32	-0.20	1.00	1996	1761
σ_ξ	0.25	0.24	0.05	0.20	0.31	1.00	1829	1962
σ_r	0.03	0.03	0.00	0.03	0.04	1.00	1976	1888
p	0.06	0.05	0.03	0.02	0.11	1.00	1790	1898
a	0.33	0.35	0.09	0.23	0.43	1.00	1216	830
μ_Y	1.29	1.37	0.64	0.45	1.93	1.01	1397	915
σ_Y	0.95	0.16	0.86	0.12	2.21	1.03	147	238

Parameter Estimates of MA models for COVID Data

Table B.4: Posterior estimates for the US data using the MA model

	Mean	MAP	Sd	10%-PI	90%-PI	split-$\hat{R}$	Bulk-ESS	Tail-ESS
β_1	0.10	0.10	0.02	0.07	0.13	1.00	2019	1872
β_2	0.15	0.15	0.02	0.12	0.18	1.00	1866	1964
β_3	0.14	0.14	0.02	0.11	0.16	1.00	1906	1749
β_4	0.19	0.20	0.02	0.17	0.22	1.00	1995	1850
β_5	0.17	0.16	0.02	0.14	0.19	1.00	2008	1748
β_6	0.08	0.08	0.02	0.06	0.11	1.00	2048	1743
β_7	0.08	0.08	0.02	0.05	0.10	1.00	2141	1921
β_8	0.05	0.05	0.02	0.03	0.08	1.00	1891	1885
β_9	0.02	0.02	0.02	0.00	0.05	1.00	2153	1961
β_{10}	0.01	0.00	0.01	0.00	0.03	1.00	1912	1984
$\beta_1^{(J)}$	0.01	0.00	0.00	0.00	0.01	1.00	1824	1679
$\beta_2^{(J)}$	0.02	0.00	0.01	0.00	0.03	1.00	1916	2005
$\beta_3^{(J)}$	0.13	0.13	0.01	0.11	0.14	1.00	1721	1925
$\beta_4^{(J)}$	0.16	0.16	0.01	0.14	0.17	1.00	2073	2046
$\beta_5^{(J)}$	0.15	0.15	0.01	0.14	0.17	1.00	2199	1780
$\beta_6^{(J)}$	0.14	0.14	0.01	0.13	0.16	1.00	2056	1924
$\beta_7^{(J)}$	0.12	0.12	0.01	0.11	0.13	1.00	2074	1846
$\beta_8^{(J)}$	0.10	0.10	0.01	0.09	0.12	1.00	2057	1751
$\beta_9^{(J)}$	0.09	0.09	0.01	0.08	0.11	1.00	1661	1886
$\beta_{10}^{(J)}$	0.09	0.09	0.01	0.07	0.10	1.00	1706	1962
d	-0.09	-0.09	0.04	-0.14	-0.05	1.00	2002	2010
σ_ξ	0.22	0.20	0.04	0.17	0.26	1.00	1586	1920
σ_r	0.02	0.02	0.00	0.02	0.02	1.00	1889	1725
p	0.09	0.05	0.05	0.03	0.15	1.00	2025	1789
b	0.50	0.48	0.14	0.33	0.68	1.01	844	900
μ_Y	1.15	1.26	0.78	0.24	2.02	1.00	1617	1887
σ_Y	1.27	0.71	0.88	0.37	2.47	1.00	525	493

Table B.5: Posterior estimates for Spain using the MA model

	Mean	MAP	Sd	10%-PI	90%-PI	split-$\hat{R}$	Bulk-ESS	Tail-ESS
β_1	0.11	0.11	0.02	0.09	0.13	1.00	2015	1884
β_2	0.16	0.16	0.02	0.14	0.18	1.00	2084	1879
β_3	0.18	0.19	0.02	0.16	0.20	1.00	1882	1852
β_4	0.15	0.15	0.02	0.13	0.18	1.00	2090	1766
β_5	0.11	0.11	0.02	0.09	0.13	1.00	2137	2018
β_6	0.04	0.04	0.02	0.02	0.06	1.00	1981	1923
β_7	0.05	0.04	0.02	0.03	0.07	1.00	1906	1924
β_8	0.06	0.06	0.02	0.04	0.08	1.00	1855	1923
β_9	0.07	0.07	0.02	0.05	0.09	1.00	2037	1881
β_{10}	0.06	0.06	0.02	0.04	0.08	1.00	1688	1886
$\beta_1^{(J)}$	0.02	0.00	0.02	0.00	0.04	1.00	2151	1989
$\beta_2^{(J)}$	0.01	0.00	0.01	0.00	0.02	1.00	2047	1899
$\beta_3^{(J)}$	0.06	0.06	0.03	0.03	0.10	1.00	1796	1833
$\beta_4^{(J)}$	0.16	0.16	0.03	0.12	0.19	1.00	2008	1882
$\beta_5^{(J)}$	0.12	0.12	0.03	0.09	0.15	1.00	1879	1924
$\beta_6^{(J)}$	0.07	0.07	0.02	0.03	0.10	1.00	1919	1947
$\beta_7^{(J)}$	0.09	0.08	0.03	0.05	0.12	1.00	1973	1942
$\beta_8^{(J)}$	0.13	0.12	0.03	0.10	0.16	1.00	2013	1967
$\beta_9^{(J)}$	0.18	0.18	0.03	0.15	0.22	1.00	1993	1938
$\beta_{10}^{(J)}$	0.17	0.16	0.03	0.13	0.20	1.00	1835	1858
d	-0.28	-0.28	0.05	-0.34	-0.21	1.00	2038	1655
σ_ξ	0.33	0.32	0.05	0.27	0.40	1.00	1910	1938
σ_r	0.03	0.03	0.00	0.03	0.04	1.00	1923	1814
p	0.05	0.03	0.03	0.01	0.09	1.00	1488	1414
b	0.21	0.20	0.09	0.10	0.32	1.00	1881	1855
μ_Y	1.19	0.19	1.06	0.17	2.52	1.00	406	663
σ_Y	1.34	0.62	1.02	0.33	2.84	1.01	174	217

Table B.6: Posterior estimates for Poland using the MA model

	Mean	MAP	Sd	10%-PI	90%-PI	split-$\hat{R}$	Bulk-ESS	Tail-ESS
β_1	0.20	0.20	0.02	0.17	0.23	1.00	1964	1964
β_2	0.20	0.20	0.02	0.17	0.23	1.00	1882	1883
β_3	0.11	0.11	0.02	0.08	0.13	1.00	2062	2006
β_4	0.09	0.09	0.02	0.07	0.12	1.00	2028	2005
β_5	0.09	0.09	0.02	0.06	0.11	1.00	2165	2093
β_6	0.08	0.08	0.02	0.05	0.10	1.00	2019	1844
β_7	0.07	0.07	0.02	0.05	0.10	1.00	2119	2050
β_8	0.05	0.05	0.02	0.02	0.07	1.00	1859	1885
β_9	0.06	0.06	0.02	0.04	0.09	1.00	2057	2007
β_{10}	0.05	0.06	0.02	0.03	0.08	1.00	2002	1981
$\beta_1^{(J)}$	0.05	0.05	0.02	0.02	0.07	1.00	1736	1661
$\beta_2^{(J)}$	0.01	0.00	0.01	0.00	0.02	1.00	1799	1709
$\beta_3^{(J)}$	0.07	0.07	0.02	0.05	0.10	1.00	2171	1900
$\beta_4^{(J)}$	0.08	0.07	0.02	0.05	0.10	1.00	1890	1886
$\beta_5^{(J)}$	0.11	0.11	0.02	0.09	0.13	1.00	1919	1819
$\beta_6^{(J)}$	0.12	0.12	0.02	0.10	0.14	1.00	1864	1927
$\beta_7^{(J)}$	0.14	0.14	0.02	0.11	0.16	1.00	1741	1851
$\beta_8^{(J)}$	0.14	0.15	0.02	0.12	0.17	1.00	1755	1988
$\beta_9^{(J)}$	0.16	0.17	0.02	0.14	0.19	1.00	2109	2141
$\beta_{10}^{(J)}$	0.13	0.13	0.02	0.10	0.15	1.00	1943	1887
d	-0.26	-0.25	0.05	-0.31	-0.20	1.00	2123	1956
σ_ξ	0.25	0.24	0.04	0.20	0.31	1.00	1918	2005
σ_r	0.03	0.03	0.00	0.03	0.04	1.00	2094	2007
p	0.06	0.04	0.03	0.02	0.11	1.00	1628	1985
b	0.50	0.48	0.12	0.37	0.64	1.00	1943	1423
μ_Y	1.22	1.28	0.57	0.50	1.71	1.01	1064	399
σ_Y	0.72	0.09	0.78	0.06	1.80	1.02	154	233

B.4 Prior parameterisation for COVID data

Table B.7: Prior parameterisation for all countries using COVID data (US, Spain and Poland).

Parameter	Prior Distribution	Hyperprior 1	Hyperprior 2
Age Parameter			
$(\beta_1, \ldots, \beta_A)$	Dirichlet$(1, \ldots, 1)$	-	-
$\left(\beta_1^{(J)}, \ldots, \beta_A^{(J)}\right)$	Dirichlet$(1, \ldots, 1)$	-	-
Time Parameter			
$\Delta\kappa_t$	$\Delta\kappa_t \overset{i.i.d.}{\sim} \mathcal{N}(d, \sigma_\xi^2)$	$d \sim \mathcal{N}(0, 5^2)$	$\sigma_\xi \sim \mathcal{N}^+(0, 2^2)$
N_t	$N_t \overset{i.i.d.}{\sim}$ Bern(p)	$p \sim$ Beta$(1, 20)$	-
Y_t	$Y_t \overset{i.i.d.}{\sim} \mathcal{N}^+(\mu_Y, \sigma_Y^2)$	$\mu_Y \sim \mathcal{N}^+(0, 4^2)$	$\sigma_Y \sim \mathcal{N}^+(0, 2^2)$
a	$a \sim \mathcal{N}^+(0, 0.4^2)$	-	-
Other Parameters			
σ_r	$\sigma_r \sim \mathcal{N}^+(0, 2^2)$	-	-

Table B.8: Prior parameterisation for the England and Wales Data

Parameter	Prior Distribution	Hyperprior 1	Hyperprior 2
Age Parameter			
$(\beta_1, \ldots, \beta_A)$	Dirichlet$(1, \ldots, 1)$	-	-
$\left(\beta_1^{(J)}, \ldots, \beta_A^{(J)}\right)$	Dirichlet$(0.5, 0.5, 0.5, 5, 5,$ $5, 5, 0.5, 0.5, 0.5)$	-	-
Time Parameter			
$\Delta\kappa_t$	$\Delta\kappa_t \overset{i.i.d.}{\sim} \mathcal{N}(d, \sigma_\xi^2)$	$d \sim \mathcal{N}(0, 5^2)$	$\sigma_\xi \sim \mathcal{N}^+(0, 2^2)$
N_t	$N_t \overset{i.i.d.}{\sim}$ Bern(p)	$p \sim$ Beta$(1, 20)$	-
Y_t	$Y_t \overset{i.i.d.}{\sim} \mathcal{N}^+(\mu_Y, \sigma_Y^2)$	$\mu_Y \sim \mathcal{N}^+(0, 5^2)$	$\sigma_Y \sim \mathcal{N}^+(0, 5^2)$
a	$a \sim$ Beta$(1, 5)$	-	-
Other Parameters			
σ_ε	$\sigma_\varepsilon \sim \mathcal{N}^+(0, 2^2)$	-	-

B.5 Overview on used samplers for own model

Table B.9: Overview on selected samplers for own model

Sampler	Parameter
AF Slice Sampler	
	$(Y_3, \ldots, Y_{T-1}, \mu_Y, \sigma_Y)$
Binary Sampler	
	$N_t \quad \forall t \in \{1, \ldots, T-1\}$
Gibbs	
	d
	p
Random Walk Metropolis	
	σ_r
	σ_ξ
Slice Sampler	
	a
HMC	
	$\left(\Delta\kappa_2, \ldots, \Delta\kappa_T, \beta_1, \ldots, \beta_A, \beta_1^{(J)}, \ldots, \beta_A^{(J)}\right)$

Chapter 3

Probabilistic Population Forecasts for Small Regions

Abstract

BACKGROUND

We consider the problem of obtaining probabilistic age-specific population forecasts for small areas or subnational regions.

OBJECTIVE

We introduce Bayesian methods suitable for obtaining reliable age-specific population forecasts for small regions using the cohort-component method.

METHODS

Our approach improves fertility forecasting by extending the Lee-Carter model with an age-region interaction term. We propose to forecast net-migration counts using skewed error terms, and introduce a Dirichlet regression to model migration age patterns as well as age proportions of fertility.

RESULTS

We run our model to produce age-specific population forecasts for a set of 13 heterogeneous regions in Bavaria, Germany. We compare our method to other standard approaches and find that it produces superior out-of-sample forecasts according to both point measures and scoring rules.

CONCLUSIONS

The results indicate that the proposed Bayesian methods offer good predictive accuracy and are suitable in obtaining precise forecasts of age-specific population for small regions.

CONTRIBUTION

We introduce a new method for the probabilistic projection of subnational population that works well and outperforms current other methods.

3.1 Introduction

Reliable age-specific population forecasts for small areas are essential for effective urban planning, resource allocation, and infrastructure development. They help local govern-

ments anticipate future needs – such as schools, healthcare services, housing, and transportation – ensuring efficient resource distribution as a community grows or changes. Despite these crucial applications, population forecasts at the regional or subnational level have historically received far less attention than those at the national level. Although methods used for national forecasts can usually be adapted for regional use, significant challenges often arise. First, migration flows are far more significant at a regional level. While these flows are less important and often neglected at a national level (Booth, 2006), at a subnational level, they are typically larger relative to the population and can become a primary driver of demographic change. Furthermore, regional migration is inherently difficult to predict and can be heavily influenced by external factors such as wars and changes in political structures, for example the dissolution of the German Democratic Republic, which resulted in significant out-migration to economically stronger regions of West Germany between 1990 and 1995 (Wolff et al., 2022). Second, available data for subnational areas are often shorter, of poorer quality, more erratic and can include zero cell counts (Wilson et al., 2022). Third, it is more difficult to predict regional populations, leading to higher forecast errors at the regional than at the national level (Tayman, 2011). Nevertheless, interest in population forecasts for small areas has grown recently. For example Cameron and Poot (2011), Swanson and Tayman (2014), and Wilson (2012) and Swanson et al. (2025) each apply different techniques to incorporate uncertainty intervals into existing deterministic regional population forecasts.

Moreover, Alexander and Alkema (2022) propose a probabilistic cohort-component method to predict the number of women of reproductive age. Their method incorporates census data and aims to minimize data requirements. Yu et al. (2023) predict subnational populations in the state of Washington by scaling national input data to obtain regional estimates. Wiśniowski and Raymer (2025) propose a multi-regional population projection model that includes inter-regional migration counts, i.e., migration flows between regions, to forecast age-specific population for Australian states. However, such data may not always be available. Moreover, their method is tailored for larger subnational regions. Further work is required to adapt it to small areas where the incidence of events is low or zero (Wiśniowski & Raymer, 2025). In this paper, we aim to address the challenge of obtaining reliable, age-specific population forecasts at a small regional level using a probabilistic implementation of the cohort-component method. To this end, we introduce new methods specifically designed to predict age-specific population at a subnational level while accounting for regional dependencies.

The paper is organized as follows: We first describe the cohort-component method and its probabilistic extension and give a brief overview on Bayesian methods in demography. Then we introduce the data that is used to create population forecasts for small areas. Next, we outline our proposed methodology for obtaining regional predictions of mortal-

ity, fertility and net-migration. We assume the same model for both sexes to calculate mortality and net-migration, although they are estimated separately. However, for notational convenience a sex-specific subscript is omitted. Lastly, we present the results of the population forecast and end the paper with a discussion.

3.1.1 Cohort-component method

The most commonly used approach for population projection are cohort-component methods. These are based on the fundamental equation, stating that the population (E) at time $t+1$ is equal to the population at time t plus the number of births (B) plus the number of in-migrants (I) minus the number of deaths (D) minus the number of out-migrants (O) (Preston et al., 2000):

$$E_{t+1} = E_t + B_t - D_t + I_t - O_t. \tag{3.1}$$

The cohort-component method of population projection uses the fundamental equation as its key underlying concept. It takes a baseline population with a specific age structure and projects it forward using age-specific mortality, fertility, and net-migration rates. The process is repeatedly applied to obtain a future population for each period after the base period and was mathematically formalized by Leslie (1945). A more detailed description can be found in Raftery and Ševčíková (2023) or Preston et al. (2000, ch. 6).

Traditional population projections based on the cohort-component method typically define deterministic scenarios based on combinations of mortality, fertility, and migration. However, this approach has several issues. First, it is impossible to define prediction intervals. Second, the components remain fixed in their respective low, medium, or high variants throughout the forecast duration, which can be unrealistic assumption (Booth, 2006). Third, a scenario based approach provides inconsistent indicators of uncertainty (Lee, 1998). To overcome these problems, probabilistic cohort-component methods have been introduced. These methods account for uncertainty by assigning a statistical model to each component, and therefore allow for a probabilistic projection of the future population. This provides a more realistic understanding of the uncertainty of the forecast. Probabilistic implementations of the cohort-component projection method are typically implemented within a Bayesian framework (e.g., Alexander & Alkema, 2022; Yu et al., 2023). While some authors distinguish between deterministic population projections and probabilistic population forecasts, we use these terms interchangeably in this work to refer to probabilistic predictions of future population size and structure.

In this paper, we propose a probabilistic implementation of the cohort-component method that is similar in style to the approach used by the United Nations (UN) for their official

population forecasts for all countries worldwide (see United Nations, 2024).

Probabilistic population forecasts can be obtained using a Monte Carlo approach, which involves generating random draws (or trajectories) from the probability distributions of the demographic components: mortality, fertility, and net-migration. Each trajectory represents a stochastic realization of how these components may evolve over time. For each realization, age-specific population values are projected forward in time. The process is repeated across all random draws, producing a range of possible future population outcomes. From these simulations, uncertainty intervals (e.g., 95% prediction intervals) can be derived by calculating empirical quantiles. For our projections, we assume that half of the net-migration occurs at the beginning and the other half at the end of each projection period. The population is projected forward for each sex separately, utilizing sex-specific inputs for mortality and net-migration. The total number of births is divided into male and female births using a sex ratio at birth (SRB) estimate derived from past data, assumed to be constant over time.

3.1.2 Bayesian Methods

Bayesian methods are becoming increasingly popular in demographic literature (Bijak & Bryant, 2016) and have been successfully applied to model subnational estimates for each input of the cohort-component method, that is, mortality, fertility, and migration rates. More precisely, numerous researchers utilize Bayesian hierarchical models to pool information across time and space to model subnational mortality using principal component analysis and singular value decomposition (Alexander et al., 2017; Dharamshi et al., 2025), the TOPLAS relational model, a spline based approach, (Rau & Schmertmann, 2020; Schmertmann & Gonzaga, 2018) or classic random effect models extended with a spatially structured parameter (Congdon, 2014; Goes, 2024). For subnational fertility, Ševčíková et al. (2018) propose Bayesian methods to produce subnational estimates of total fertility rate, while Schmertmann et al. (2013) use empirical Bayesian methods to smooth regional fertility data before applying a Brass relational model to obtain subnational fertility estimates. Bryant and Zhang (2016) introduce a Bayesian method for obtaining subnational migration estimates, and Yu et al. (2023) adapt a Bayesian method by Azose and Raftery (2015), originally designed for national-level data, to generate subnational forecasts.

In this paper, we propose a method employing a Bayesian implementation of the cohort-component method to generate probabilistic forecasts of age-specific populations. Here, we introduce new methods to predict each input of the cohort-component method that are specifically suited for subnational data. More precisely we propose skewed error terms that allow us to predict net-migration counts more accurately which are heavily influenced

by external factors. In addition, we introduce a novel approach to forecast age-specific fertility rates using a Dirichlet regression. We evaluate our proposed methods on out-of-sample data and compare the results to the standard approaches in the demographic literature.

The paper is organized as follows: We first describe the data that is used to create population forecasts for small areas. Next, we outline our proposed methodology for obtaining regional predictions of mortality, fertility and net-migration. We assume the same model for both sexes to calculate mortality and net-migration, although they are estimated separately. However, for notational convenience a sex-specific subscript is omitted. Lastly, we present the results of the population forecast and end the paper with a discussion.

3.2 Data

We aim to forecast population counts for small areas or regions. There is no universally accepted definition for what constitutes a "small area". For example, Wilson et al. (2022) use the term for areas with populations under 100,000, though they acknowledge this cutoff is somewhat arbitrary. The European Union defines a Nomenclature of territorial units for statistics (NUTS), differentiating various region sizes[1]. Their smallest unit, NUTS-3, typically denotes a region within a state or province. For example, in Germany, NUTS-3 regions are often districts (Landkreise) or independent cities (kreisfreie Städte). The exact definition varies from country to country, but the general principle is that they represent relatively small, geographically defined areas. In the following, we will use the term small areas to refer to NUTS-3 regions[2].

To illustrate our method, we apply it to a specific set of NUTS-3 regions, within the state of Bavaria, Germany, called Upper Franconia. Though in general, the method that is introduced can be broadly applied to any set of NUTS-3 regions, or even lower-level areas that the European Union refers to as local administrative units (LAU). Bavaria is the largest state of Germany in terms of area and is located in the south east of the country. Upper Franconia is one of seven administrative subunits within Bavaria, located in the north east and consists of $R = 13$ NUTS-3 regions. We believe these to be a typical set of small areas in a sense that they are heterogeneous in age structure, GDP per capita (cf. Fig. C.2 in the appendix), as well as population size population size, ranging from around 40,000 to 150,000. In addition, several of the regions have populations with a high share of students with a distinct age profile, leading to a specific fertility schedule

[1] see https://ec.europa.eu/eurostat/web/nuts, accessed 17.07.25

[2] This is a different to the definition used in the small area community, see e.g. Rao and Molina (2015).

different from that of other regions. Moreover, due to their differences in GDP, and thus economic activity, some regions experience vastly different net-migration.

The data for evaluating our method are provided by the Statistical Office of Bavaria (Bayerisches Landesamt für Statistik; BLfS). The data are publicly available and can be downloaded from GENESIS, the database of the Bavarian statistical institute. The datasets consist of population counts (BLfS 2025), counts of deaths (BLfS 2022b), counts of births (BLfS 2024a) and counts of in and out-migration (BLfS 2024b). The death and part of the migration data are disaggregated by age grouped into five-year intervals, starting from $[0, 5)$ up to $[80, 85)$ with an open-ended age group of 85+, resulting in a total of $X = 18$ age groups. The counts of births are also provided in five-year age groups, with an open-ended age group for the youngest (< 20) and the oldest (40+). However, these datasets are available for different time periods. Specifically, counts of births are available from 1995 – 2023, while counts of deaths are only available from 2000 – 2017. Unfortunately, after 2017, counts of deaths at regional level are no longer available in a five-year age grid due to a change in data protection regulations. Thus, to estimate mortality rates we use data from 2000–2017, while estimates of fertility can be obtained using a longer time series. Age-specific population counts are available from 2000 – 2024.

For migration analysis, two datasets are available. The Statistical Office of Bavaria provides migration counts, distinguishing between in-migration and out-migration, using a broad age grid ($<$ 18, 18-25, 25-30, 30-50, 50-65, 65+). These data cover the period from 2000 to 2023 and are used to predict future total migration counts. However, this age grid is too coarse for calculating the migration schedule (see Section Section 3.5). Therefore, we contacted the statistical institute to request migration data on a more granular five-year age grid. They kindly provided the necessary data, though it is only available for the years 2011 to 2023. We used these finer-grained counts to calculate the age-specific migration schedule. Notably, this dataset contains eight missing data points out of a total of 3042 for both males and females. The missing values do not follow a discernible pattern. Instead of performing a complete case analysis, we chose to impute the missing data points. Given the small number of missing values, we opted for single imputation rather than multiple imputation, assuming ignorability – that is, the distribution of the data is the same for both observed and unobserved data points. Missing values were imputed using the `mice` package (multivariate imputations by chained equations) in `R` (van Buuren & Groothuis-Oudshoorn, 2011), specifically employing the CART (classification and regression trees) method.

3.3 Mortality

To produce forecasts of mortality rates we follow the ideas of Goes (2024) using a Renshaw-Haberman (RH) model (Renshaw & Haberman, 2006) extended with a spatially structured regional effect. This is done to capture correlation patterns across regions and thereby increase the model fit and predictive accuracy. Let $D_{x,t,r}$ denote the number of deaths and $E_{x,t,r}$ the population at risk or exposure in age group $x \in \{1, \dots, X\}$, year $t \in \{1, \dots, T\}$ and region $r \in \{1, \dots, R\}$. For the number of deaths, we assume a Poisson likelihood with $D_{x,t,r} \sim \text{Poi}(m_{x,t,r} \cdot E_{x,t,r})$, where $m_{x,t,r}$ denotes the underlying mortality rate, with

$$\log(m_{x,t,r}) = \alpha_x + \beta_x^{(1)}\kappa_t + \beta_x^{(2)}\gamma_k + \nu_r + \varepsilon_{x,t,r}$$

and $\varepsilon_{x,t,r} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$. Here, the parameter α_x denotes the average log-mortality rate at age x and models the general shape of mortality by age. The parameter κ_t estimates the global change over time and γ_k, the global effect of cohort k. The index k is a function of age and period. If the intervals are of different lengths, that is, if the age intervals are M times wider than the period intervals, the cohort index is given by $k = M(X-x)+t$ with $k \in \{1, \dots, M(X-1)+R\}$ (Heuer, 1997). The parameters $\beta_x^{(1)}$ and $\beta_x^{(2)}$ measure the response to changes of κ_t and γ_k, respectively, at age x. The regional effect ν_r captures spatial dependencies which is modeled using a BYM2 prior proposed by Riebler et al. (2016), an extension of the famous Besag-York-Mollié (BYM) model (Besag et al., 1991). An error term $\varepsilon_{x,t,r}$ accounts for overdispersion. To make the model identifiable we invoke the typical constraints, that is, $\sum_x \beta_x^{(1)} = \sum_x \beta_x^{(2)} = 1$ and $\sum_t \kappa_t = \sum_k \gamma_k = \sum_r \nu_r = 0$.

3.3.1 Out-of-sample evaluation

To demonstrate the validity of our proposed model, we evaluate its predictive performance against a classical benchmark: the standard Lee-Carter model (Lee & Carter, 1992), estimated separately for each region. Following the approach of Goes (2024), we fit both models using data from 2000 to 2014. We then generate out-of-sample predictions of future deaths for the years 2015 to 2017 and compare their predictive accuracy against observed values. To obtain future deaths for both models, we forecast age-specific mortality and multiply those forecasts by the observed values of exposure. The resulting value is then used as the parameter of a Poisson distribution from which future deaths are drawn. We denote our proposed model as RH_BYM2 and the standard Lee-Carter model as LC.

We calculate the mean absolute error (MAE), the root mean squared error (RMSE), and the coverage. We refrain from calculating the very popular mean absolute percentage error (MAPE), due to the high amount of observed zero deaths in our data set. The coverage of a prediction interval is defined as the proportion of time that the true value

lies within this interval. Ideally, coverage values should be close to its nominal value, which in this case is 0.95. Furthermore, using Bayesian methods allows us to obtain not just point forecasts but an entire predictive distribution, which can be compared using so called scoring rules. To compare these predictive distributions, we utilize the log score (LogS) and ranked probability score (RPS). For a detailed explanation, readers are referred to Gneiting and Raftery (2007) and Czado et al. (2009). Similar to point measures, lower scores indicate a better fit. These metrics can be computed in `R` based on samples from the posterior predictive distribution using the *scoringRules* package (Jordan et al., 2019).

The results demonstrate that our proposed model provides superior forecasts of age-specific mortality across all metrics. Specifically, it achieves lower values for both point measures (MAE and RMSE) and scoring rules, for both men and women. Additionally, the coverage values are closer to the nominal value of 0.95. The results are shown in Table 3.1.

Table 3.1: Out-of-sample performance of mortality forecasts

Model	**MAE**	**RMSE**	**Coverage**	**LogS** $\cdot 10^{-2}$	**RPS** $\cdot 10^{-2}$
Men					
RH_BYM2	**3.47**	**6.27**	**0.97**	**16.65**	**17.42**
LC	4.09	7.28	0.98	17.42	19.92
Women					
RH_BYM2	**3.61**	**8.00**	**0.97**	**15.22**	**18.12**
LC	4.03	8.55	0.97	16.20	20.32

Note: Values in bold denote the best of the column. MAE = mean absolute error. RMSE = root mean squared error. Coverage denotes the coverage for a nominal value of 0.95. LogS = log score. RPS = ranked probability score

3.4 Fertility

Let $B_{x,t,r}$ denote the number of live births that women have in age group x, year t and region r. Moreover, let $E^{(F)}_{x,t,r}$ denote the female population at risk in age group x, year t and region r. We assume that age-specific fertility rates $f_{x,t,r}$ are positive for age groups corresponding to the age interval $[15, 45)$ and zero otherwise. This is due to the following reason. The data of births is available in five-year age groups, except for the lowest (< 20) and highest ($40+$) age groups. To generate future births, we need to multiply the age-specific fertility rate with the respective age-specific female population. Therefore, we assume that all of births of the lowest age group occur within the interval of $[15, 20)$ and all of the births of the highest age-group within $[40, 45)$, resulting in positive age-specific fertility rates for the age interval of $[15, 45)$.

The total fertility rate (TFR) is defined as $TFR_{t,r} = \sum_x 5 \cdot f_{x,t,r}$. In addition to the

TFR, we can define the age proportion:

$$\phi_{x,t,r} = \frac{5 \cdot f_{x,t,r}}{TFR_{t,r}}.$$

This represents the fraction or proportion of total fertility occurring in age group x. Note that $\sum_x \phi_{x,t,r} = 1$ for all r and t. There are multiple ways to model and later on forecast fertility rates. One can either model them directly or indirectly. A typical model for direct estimation of age-specific fertility rates is the Lee-Carter model, originally designed for mortality rates but shortly thereafter adapted for age-specific fertility rates (Lee, 1993). Alternatively, one can use the UN's approach, where estimates of TFR and age-proportion are backtransformed into age-specific fertility rates (see Ševčíková et al. (2016) for details). We refer to the former as the direct approach and the latter as the indirect approach. Bohk-Ewald et al. (2018) compare various methods to forecast cohort fertility rates and found both the direct approach using a Lee-Carter model and the indirect approach to be among the best-performing methods.

3.4.1 Direct estimation

In the direct approach, the age-specific fertility rate $f_{x,t,r}$ is estimated using an extension of the Lee-Carter model with a Poisson likelihood for the total number of births, thus $B_{x,t,r} \sim \text{Poi}\left(f_{x,t,r} \cdot E^{(F)}_{x,t,r}\right)$ with

$$\log(f_{x,t,r}) = \alpha_x + \beta_x \kappa_t + \delta_{x,r} + \varepsilon_{x,t,r}, \tag{3.2}$$

where $\varepsilon_{x,t,r} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\varepsilon)$. The parameter α_x denotes an age-specific intercept and can be thought of as the initial fertility level, while the parameter κ_t represents the change in fertility over time scaled by an age-specific factor β_x. The model of Eq. (3.2) is different from the standard Lee-Carter model due to the addition of an age-region interaction $\delta_{x,r}$. This effect is included because varying fertility patterns were observed for different regions. Instead of the above interaction term, one could also think of naturally extending the Lee-Carter model by setting $\delta_{x,r} = \beta^{(2)}_x \nu_r$, where $\boldsymbol{\beta}^{(2)} = \left(\beta^{(2)}_1, \ldots, \beta^{(2)}_X\right)^\top$ denotes another vector of age effects and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_R)^\top$ represents a vector of regional effects. However, this approach introduces an additional $X + R - 2$ parameters (subject to two constraints), while the more general $\Delta \in \mathbb{R}^{X \times R}$ with entries $\delta_{x,r}$ introduces a total of $X \cdot R - X$ parameters (subject a corner constraint, details are given below). Since the latter approach introduces more parameters into the model, it is more flexible. Moreover, we compared both the in- and out-of-sample fit of both models and found the approach given in Eq. (3.2) to be superior. Details are omitted for brevity.

Using an age-region interaction term is different from the approach presented in Sec-

tion 3.3, where the RH model was extended by the addition of a spatially structured regional effect. It would have also been possible to extend the Lee-Carter model of Eq. (3.2) with the same effect, that is, to set $\delta_{x,r} = s_r$, though we chose the interaction term for the same reason as stated above. Both the in-sample and out-of-sample performance of Eq. (3.2) was better than that of a model with a spatially structured effect. However, the flexible approach using the interaction term is only possible due to the relatively few age-groups and regions within our dataset. For more age groups (e.g. in the mortality model) and/or more regions, the number of parameters to estimate grows substantially and thus increases computational burden. In such a case, a Lee-Carter model extended with a spatially structured effect may be preferred.

To ensure model identifiability of Eq. (3.2), we impose the following constraints: $\kappa_1 = 0$, $\|\boldsymbol{\beta}\|_2^2 = \sum_{x=1}^{X} \beta_x^2 = 1$, $\kappa_t \geq 0$ for all t, and $\delta_{x,R} = 0$ for all x, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_X)^\top$ denotes the vector of age parameters. Note, that $\boldsymbol{\beta}$ is constrained to have a norm of one, diverging from the more common sum-to-one constraint. This choice was made primarily for computational reasons, as we observed this specification to be more stable in `Stan`, the software we use for parameter estimation, than the more common sum-to-one constraint. However, this led to us having to set the additional constraint of $\kappa_t \geq 0$ (see Section 3.6 for details).

3.4.2 Indirect Estimation

The indirect approach forecasts both TFR and age proportions and then converts them into age-specific fertility rates using

$$f_{x,t,r} = \frac{\phi_{x,t,r} \cdot \mathrm{TFR}_{t,r}}{5}.$$

This method is implemented by the UN for their world population forecasts (Ševčíková et al., 2016). Thus, we need to obtain regional estimates of TFR and age-proportions to then backtransform these into age-specific fertility rates.

Ševčíková et al. (2018) propose a method to predict subnational TFRs. Similar to them, we assume that the regional TFR follows an Bayesian hierarchical autoregressive model of order one, AR(1). More precisely,

$$\mathrm{TFR}_{t,r} - \mu_r = a_r \cdot (\mathrm{TFR}_{t-1,r} - \mu_r) + \omega_{t,r}, \tag{3.3}$$

with $\omega_{t,r} \overset{iid}{\sim} \mathcal{N}(0, \sigma_r^2)$, a region specific intercept μ_r and autoregressive parameter a_r. Both the intercept and autoregressive parameters are assigned hierarchical priors (see Section C.6 for an overview on the selected priors). In Eq. (3.3), region-specific TFRs are modeled directly, in contrast to Ševčíková et al. (2018), where subnational TFR estimates

are obtained by scaling country-specific TFR estimates using a regional scaling factor. This scaling factor is proposed to follow an AR(1) model with a mean of one, implying that in the long term, forecasts of subnational TFRs converge to the respective TFR forecast of the country. For some regions, such as those with a significant population of students, this is an unrealistic assumption. To circumvent this issue, Yu et al. (2023) adjust the TFR data for student populations by scaling past TFR values. However, this requires additional data, and it is unclear if this adjustment performs reliably. We therefore model the regional TFRs directly, assuming a distinct mean for each region.

To forecast age proportions, Ševčíková et al. (2016) propose a linear combination of the past observed trend and the past global trend of all age proportions in the dataset. Again, this method seems unsuitable for regions with college populations, as these areas may exhibit distinct behavior compared to other regions. Moreover, their method is deterministic. That is, they do not specify a probability distribution for the age proportion, making it impossible to generate probabilistic forecasts for $\phi_{x,t,r}$. To circumvent these issues, we propose to model the regional age proportions using a Dirichlet regression (see e.g. Gueorguieva et al. (2008)), given their inherent constraints of $\sum_x \phi_{x,t,r} = 1$ and $\phi_{x,t,r} \in [0, 1]$. Let $\boldsymbol{\phi}_{t,r} = (\phi_{1,t,r}, \ldots, \phi_{X,t,r})^\top$ denote a multivariate random variable with $X = 6$ groups, then $\boldsymbol{\phi}_{t,r} \sim \text{Dirichlet}(\psi_{1,t,r}, \ldots, \psi_{X,t,r})$, where

$$\log(\psi_{x,t,r}) = \alpha_x + \beta_x \kappa_t + \delta_{x,r}.$$

The linear predictor $\log(\psi_{x,t,r})$ has the same structure as that of a Lee-Carter model extended with an age-region interaction $\delta_{x,r}$. To make the model identifiable we impose the same constraints as for Eq. (3.2), that is, $\kappa_1 = 0$, $\|\boldsymbol{\beta}\|_2^2 = \sum_{x=1}^{X} \beta_x^2 = 1, \kappa_t \geq 0$ for all t and $\delta_{x,R} = 0$ for all x. An explanation for their choice is given in Section 3.6. We tested multiple specifications for the linear predictor of $\psi_{x,t,r}$, such as a random intercept random slope model with the inclusion of an age-region interaction effect. Nonetheless, the Lee-Carter structure performed best according to the Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2010) as well as out-of-sample criteria. Results are omitted for brevity. To the best of our knowledge, this is the first attempt to estimate age proportions using a Dirichlet regression.

3.4.3 Out-of-sample comparison

We now assess the predictive performance of both approaches to ultimately decide which to use in the population projection. We fit both approaches on a subset of the data, specifically for the years 1995–2015, and compare their predictive accuracy against the ground truth, which is the observed number of births between 2016 and 2023. Forecasts for births are obtained as follows: For the direct approach, forecasts of age-specific fertil-

ity rates are multiplied by known exposure values. We then draw samples from a Poisson distribution, using the product of age-specific fertility and exposure as the distribution's parameter. For the indirect approach, forecasts of age-proportion and TFR are transformed to obtain forecasts of age-specific fertility rates, which are then multiplied by known exposure values to yield birth forecasts. Additionally, we compare the predictive accuracy of our models with the standard Lee-Carter model, estimated for each region separately, a method implemented by Myrskylä et al. (2013) for predicting cohort fertility rates across multiple countries, showing good predictive performance. As with the evaluation of mortality rates, we calculate the MAE, RMSE, coverage, and both scoring rules for fertility forecasts. Additionally, we include the mean absolute percentage error (MAPE), since there are no observed zero births for any specific age group in our dataset.

Examining the out-of-sample performance of all models, we find that the direct approach – more specifically, the Lee-Carter extension with an age-region interaction term described in Eq. (3.2) – is superior in the most categories, albeit by a small margin. This is followed by the indirect approach, and then the Lee-Carter approach (see Table 3.2). The mean forecast of the direct and indirect approach are very close to each other, with the former being marginally better in terms of MAE and RMSE, while the latter is slightly superior according to the MAPE. The standard Lee-Carter method performs worse based on the point measures. Furthermore, the direct and indirect approaches have coverages of 0.9, which is close to the nominal value of 0.95, and thus they effectively capture the underlying uncertainty. However, the standard Lee-Carter model exhibits the highest coverage values. This can be explained by the fact that the Lee-Carter method allows the error term for each region to have a different variance. In contrast, the error term in the direct approach has the same variance for all regions. Looking at the scoring rules, we find the predictive distributions of the two proposed approaches outperform the standard Lee-Carter model. According to the LogS, the indirect approach performs slightly better, while the direct approach is superior in terms of the RPS. However, the calculation of the LogS based on samples from the posterior predictive distribution is highly variable, whereas the variability is lower for the RPS, when calculated using the empirical posterior predictive cumulative distribution function. Therefore, Krüger et al. (2021) recommend the use of the RPS when estimating the scoring rules based on draws from the posterior predictive. Consequently, we place more value on the RPS results and choose to use the direct approach to generate our forecasts of age-specific fertility rates, which will serve as an input for the cohort-component method. Nonetheless, both proposed methods perform better than the classical Lee-Carter model.

Table 3.2: Out-of-sample performance of the three fertility approaches

Model	MAE	RMSE	MAPE	Coverage	LogS $\cdot 10^{-3}$	RPS $\cdot 10^{-3}$
Direct approach	**14.76**	**22.28**	18.13	0.90	2.59	**6.72**
Indirect approach	14.85	22.89	**17.45**	0.90	**2.57**	6.78
Lee-Carter	17.74	26.80	22.98	**0.92**	2.64	7.58

Note: Values in bold denote the best of the column. MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95. LogS = log score. RPS = ranked probability score

3.5 Migration

In Section 3.1 we mentioned that forecasts of net-migration by age are needed to produce regional population forecasts. Often, net-migration totals are distributed across age to obtain age-specific net-migration rates. However this does not take into account the different age-distributions of in- and out-migration and net-migration rates do not have a consistent age pattern, which give bias results (Rogers, 1990). Therefore, to estimate subnational migration, we broadly follow the approach of Ševčíková et al. (2024). They decompose forecasts of total net-migration counts into future in- and out-migration counts, which are subsequently transformed into age-specific in- and out-migration counts. In general, instead of counts, one can also model net-migration rates and then transform these into age-specific in- and out-migration counts (see Welch et al., 2024). However, as data for net-migration counts are available, we opted for the former approach.

Germany experienced a steady increase in net-migration in the early 2010s, peaking in 2015 with a net total of 1,139,402 according to the Federal Office for Migration and Refugees (FOMR 2016), which is a 50% increase compared to the previous year, with Syria being the primary country of origin due to the Syrian civil war. The full scale invasion of Russia against Ukraine in 2022 triggered another refugee movement in Europe – the biggest since World War II – resulting in Germany experiencing an increase in net-migration totals of over 400% compared to the year prior, reaching a net total of 1,462,089 according to the Federal Ministry of Interior, Build and Community (FMI) and the FOMR (FMI and FOMR 2024). These two events can be considered as two one-period migration shocks, that need to be accounted for when modeling net-migration counts for regions in Germany or more broadly across Europe. As expected, both effects can be seen when plotting net-migration counts of our regions in Upper Franconia with pronounced peaks in 2015/2016 and 2022 (see Fig. 3.1). Furthermore, we observe that the migration shock related to the Syrian civil war was lagged by a year for a single region, specifically the city of Bamberg. This is due to the so called AnkER facilities (standing for arrival, decision, and return facility), which were established in parts of Germany in 2016 to centralize asylum application processes before distributing migrants to other areas of the country. The only AnkER facility in Upper Franconia is situated in

the city of Bamberg (FOMR 2021).

Figure 3.1: Net-migration counts for all regions in Upper Franconia for both females and males

Note: Each of the 13 regions is colored differently. Dashed lines in 2015 and 2022 denote years with unusually high net-migration totals in Germany.

To model total net-migration rates for multiple regions, Azose and Raftery (2015) propose a hierarchical AR(1) approach using Gaussian error terms, which can also be used to model total net-migration counts. However, such error terms are not ideal for handling past migration shocks, as these events primarily manifest as positive outliers. A normal or t-distribution is, however, symmetric, meaning positive and negative migration shocks are equally likely. This is not reflected in the data. Fernández and Steel (1998) introduce a general class of skewed error terms, which can be included into the autoregressive modeling equation of Azose and Raftery (2015). This allows for explicitly modeling positive outliers and therefore better to describe the observed data while also allowing for the possibility of a future one-time shock event. Consider a univariate random variable with a probability density function $f(\cdot)$ that is continuous, unimodal, and symmetric around 0 (e.g., a normal or a t-distribution). For ease of notation, we assume the scale parameter of $f(\cdot)$ to be one, though in theory the scale parameter of $f(\cdot)$ can be any positive value. An additional parameter $\gamma \in (0, \infty)$ accounts for the degree of skewness, generating the following class of skewed distributions:

$$p(\varepsilon|\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \left[f\left(\frac{\varepsilon}{\gamma}\right) \mathbb{1}_{[0,\infty)}(\varepsilon) + f\left(\varepsilon\gamma\right) \mathbb{1}_{(-\infty,0)}(\varepsilon) \right]. \tag{3.4}$$

Here, $\mathbb{1}_{[0,\infty)}(\varepsilon)$ denotes an indicator function that equals one if ε lies in the interval $[0, \infty)$ and zero otherwise. The parameter γ accounts for the degree of skewness, with $\gamma = 1$ indicating no skewness, while $\gamma \in (0, 1)$ indicates left and $\gamma \in (1, \infty)$ right skewness. For further details, we refer to Fernández and Steel (1998). We can use Eq. (3.4) to introduce skewed normal error terms with a heavy right tail, allowing for a higher probability of positive outliers, that is, positive mortality shocks.

To model total net-migration counts we employ the AR(1) model of Azose and Raftery (2015) with a skewed normal error term given by

$$N_{t,r} - \mu_r = a_r \cdot (N_{t-1,r} - \mu_r) + \varepsilon_{t,r} \cdot \sigma_r, \tag{3.5}$$

where $\varepsilon_{t,r}$ follows a skewed normal distribution given in Eq. (3.4), scaled by a region specific variance parameter σ_r. Instead of using skewed errors, we could also follow the ideas of Goes et al. (2025) and introduce a global migration shock parameter, where for a given time period $\tilde{t}$ the migration counts for all regions are simultaneously shifted upwards. However, as evident in Fig. 3.1, not all regions experienced the migration shock of the Syrian civil war in 2015 at the same time, making the method less suitable for the data at hand.

3.5.1 Out-of-Sample validation

To validate our methodology, that is, to see if the skewed normal error terms of Eq. (3.4) provide better out-of-sample forecasts compared with the Gaussian errors, we compare the predictive accuracy of both approaches in conjunction with an AR(1) structure. We estimate the model of Eq. (3.5) with both skewed normal errors as well as Gaussian errors. The former we call Azoze-Raftery skewed while the latter is simply denoted as Azoze-Raftery model. To generate forecasts using Eq. (3.5), we need the ability to draw random samples from Eq. (3.4). This is achieved by generating random draws from a uniform distribution $u \sim \text{Unif}(0, 1)$ and plugging those draws into the quantile function of Eq. (3.4). Derivation of the distribution as well as the quantile function of Eq. (3.4) can be found in the appendix.

The models were fitted on data from 2000–2018 and future net-migration counts for 2019–2023 were forecasted to assess their predictive accuracy. A different out-of-sample window was specifically chosen compared to that in Section 3.4 to allow the data to return to a more typical level after the high net-migration values observed in 2015–2017 (cf. Fig. 3.1). After generating forecasts of net-migration counts for both models, these were compared against the true observed counts. As before, we analyzed the predictive accuracy of the posterior mean using classical point measures and also compared the entire predictive distribution using scoring rules. For men, we observe a comparable performance in terms

of point measures, which is unsurprising given that both models share the same underlying structure. However, when examining the coverage and scoring rules, the skewed error term proved more effective at describing the predictive distribution, indicated by lower scores and broader coverage. For women, we observe similar findings: Both approaches yield comparable point forecasts, yet, the skewed error terms provide a superior description of future migration shocks. The detailed results are presented in Table 3.3.

Table 3.3: Out-of-sample performance of net-migration forecasts

Model	MAE	RMSE	MAPE	Coverage	LogS	RPS $\cdot 10^{-3}$
Men						
Azoze-Raftery skewed	**272.45**	**362.21**	150.37	**0.78**	**485.07**	**13.51**
Azoze-Raftery	280.98	378.13	**149.45**	0.71	493.92	13.88
Women						
Azoze-Raftery skewed	228.55	329.64	90.67	0.85	**475.81**	**11.52**
Azoze-Raftery	230.14	331.79	**89.23**	0.85	488.54	11.68

Note: Values in bold denote the best of the column. MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95. LogS = log score. RPS = ranked probability score

3.5.2 Age-specific migration

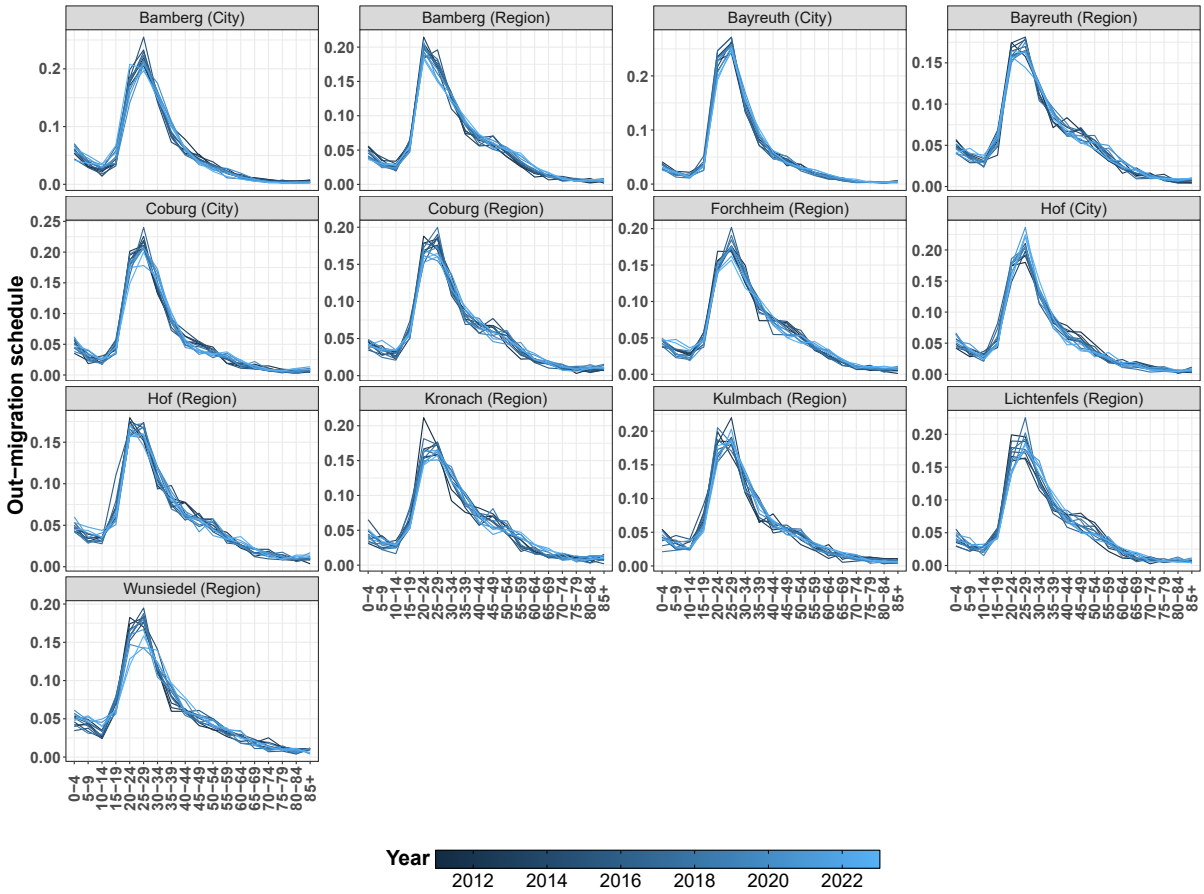
As mentioned previously, the total migration counts have to be transformed into age-specific counts. This is achieved by multiplying them by a so called migration schedule, which is defined as the age distribution of migration. Looking at the age-specific in- and out-migration counts from 2011–2023, we notice that the amount may change, yet the age distribution or pattern stays fairly consistent. The age distribution of the out-migration counts for a specific region is calculated by $R^{(out)}_{x,t,r} = O_{x,t,r} / \sum_{x=1}^{X} O_{x,t,r}$, where $\sum_x R^{(out)}_{x,t,r} = 1$ by construction. The age distribution of the in-migration counts is calculated similarly. A graphical representation of the age distribution of the male out-migration schedule, demonstrating a time-invariant pattern, is shown in Fig. 3.2. The same time-invariant pattern can also be observed for females and is shown in the appendix in Fig. C.1.

The concept of time-constant migration schedules is known in the literature and was first mathematically formalized by Rogers and Castro (1981), who introduced a statistical model to estimate this time-independent migration schedule.

Hence, using such a model, we can multiply migration totals by time-invariant migration schedules to obtain age-specific migration counts.

Ševčíková et al. (2024) propose a method to decompose total net-migration counts, $N_{r,t}$, into in-, $I_{r,t}$, and out-migration counts, $O_{r,t}$, using a mixed-effects model, which we will employ as well. These counts are then transformed into age-specific in- and age-specific

Figure 3.2: Out-migration schedule over time of men for all regions ins Upper Franconia.



out-migration counts by multiplying the total in- and out-migration counts with estimated migration schedules. More specifically, we calculate by sex

$$
\begin{aligned}
I_{t,r} &= \beta_r^{(0)} + \beta_1 N_{t,r} + \varepsilon_{t,r} \\
o_{x,t,r} &= O_{t,r} \cdot \hat{R}_{x,t,r}^{(Out)} \\
i_{x,t,r} &= I_{t,r} \cdot \hat{R}_{x,t,r}^{(In)},
\end{aligned}
$$

where $O_{t,r} = I_{t,r} - N_{t,r}$, $\varepsilon_{t,r} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{Mixed})$ and $\beta_r^{(0)}$ denotes a region specific intercept and β_1 a slope parameter. Here, $o_{x,t,r}$ and $i_{x,t,r}$ denote the age-specific out- and in-migration counts, respectively. Moreover, let $\hat{R}_{x,t,r}^{(Out)}$ and $\hat{R}_{x,t,r}^{(In)}$ denote the estimated migration schedules derived from the data, which are assumed to be constant over time. The most well-known model for estimating migration schedules is the so-called Rogers-Castro curve, proposed by Rogers and Castro (1981), with a Bayesian implementation by Yeung et al. (2023). While staying within the Bayesian framework, we attempted to implement Yeung et al. (2023)'s model using multiple time periods as input data. However, we encountered problems with convergence in `Stan` (see Section 3.6). We therefore choose to model

the age-migration schedule using a Dirichlet regression, due to its implied sum-to-one constraint. Let $R_{x,t,r}$ denote the age pattern of migration, with the index for out- or in-migration dropped for notational convenience.

We assume $\boldsymbol{R}_{t,r} = (R_{1,t,r}, \ldots, R_{X,t,r})^\top$ to be Dirichlet distributed, thus

$$\boldsymbol{R}_{t,r} \overset{iid}{\sim} \text{Dirichlet}(\psi_{1,r}, \ldots, \psi_{X,r}).$$

This implies that each region has its own set of age-specific parameters that remain constant over time. Using the posterior predictive distribution we can generate samples of both $\hat{R}^{(Out)}_{x,t,r}$ and $\hat{R}^{(In)}_{x,t,r}$.

We compare the performance of the Bayesian implementation of the Rogers-Castro model for a single time period with our Dirichlet model. We find their performance to be similar, with our Dirichlet model even showing better in-sample fit. Further details are provided in the appendix.

3.6 Estimation

The models are fitted in a Bayesian framework using the probabilistic modelling software `Stan` (Stan Development Team, 2024c), a tool for Hamiltonian Monte Carlo (HMC). Stan can be accessed through the interface `rstan` in `R` (Stan Development Team, 2024a). Built-in diagnostic measures are used to check for convergence, which are described in Vehtari et al. (2021). The parameters of all models are given weakly informative priors. Details regarding the specific choice can be found in the appendix. It should be noted, that posterior predictive checks revealed a good fit of all models to the data. For the sake of brevity the results are omitted. The code and data for all models is available on GitHub[3].

As mentioned in Section 3.4, we impose the vector of age parameters $\boldsymbol{\beta}$ in Eq. (3.2) to have a norm of one. However, this does not lead to a unique solution. More precisely, two sets of parameterizations for $\boldsymbol{\beta}$, which we denote $\mathcal{B}$ and $-\mathcal{B}$, yield the same norm, resulting in a non-unique set of parameters. We therefore have to additionally impose that $k_t \geq 0$ for all t. Despite this, we occasionally encountered convergence issues in `Stan` where estimates of $\kappa_1, \ldots, \kappa_T$ would behave differently in one chain than in another. To mitigate this, we implement another set of constraints and impose that $k_1 < k_{T/2} < k_T$, which alleviated the problem. Still, these constraints are solely due to computational reasons and might not have to be imposed using a different estimation scheme.

In addition, it is important to note that the age pattern of migration introduced in Sec-

[3]https://github.com/goesj/Population-Forecast

tion 3.5 can theoretically take on values of zero and one, i.e., $R_{x,t,r} \in [0, 1]$. However, the probability density function of the Dirichlet distribution is zero if any of its components, that is any of $R_{x,t,r}$, is zero. In addition, `Stan` works with log posteriors which are undefined at zero. Therefore, the components must satisfy $R_{x,t,r} \in (0, 1)$ Stan Development Team (see 2024b, ch. 26.1). In our data, the in-migration count for males was zero for a single age group in a single region and single year, leading to a value of 0 in the migration schedule, which we will denote $R_{x^*,t^*,r^*} = 0$. To circumvent this issue, we added a small constant and set this specific value to $R_{x^*,t^*,r^*} = 1 \times 10^{-6}$. Afterwards, we rescaled the migration schedule so that it sums to one for this specific year and region.

The computation time for each input – mortality, fertility, and net migration – varies with model complexity and the number of parameters. More specifically, due to the interaction terms, the fertility models take the longest, while the relatively simple AR(1) model has a shorter runtime. Despite these differences, `Stan`'s parallel processing capabilities ensure that the overall runtime remains moderate: the migration model completes in approximately one minute, while the fertility model requires up to 60 minutes on an Intel i5-8365U CPU (1.60 GHz) with 16 GB of RAM. However, we anticipate that computation time will increase with larger datasets, such as those including more regions or years.

3.7 Population Forecasts

Before running the full model to obtain population forecasts for all 13 regions in Upper Franconia, we first want to validate our method in the short term. Following this validation, we will present the full population forecasts for all regions from 2024–2044.

3.7.1 Out-of-sample validation

To validate our methodology, we conducted an out-of-sample population forecast and compared the predicted population with the observed one. Since the data of the population is given in age-groups of five, our projection period is of the same length. Thus, we can only project the population one projection interval, i.e., five years, at a time. We use the observed data until 2019 and forecast the population for a single projection period, from 2020 to 2024. For this, we estimate the input parameters as follows: The mortality model is trained from 2000–2017 (due to a lack of available data, see Section 3.2) and then used to forecast for the years 2018–2024. Fertility and net-migration are estimated from 1995–2019 and 2000–2019, respectively, with out-of-sample predictions generated for 2020–2024. Afterwards, we compare the predicted population for each age group and region to the observed values. Table 3.4 presents the validation results, including the RMSE, MAE, and coverage.

We observe a coverage value of 0.85, which is consistent with the coverage values of

Table 3.4: Out-of-sample validation of the age and region-specific population forecast

MAE	RMSE	MAPE	Coverage
115.46	207.22	2.83	0.85

Note: MAE = mean absolute error. RMSE = Root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95

the out-of-sample forecasts for net-migration and fertility. Coverages for out-of-sample forecasts are typically smaller than their nominal value due to the increased uncertainty inherent in out-of-sample predictions. However, this coverage remains within a satisfactory range and is comparable to population forecasts by other authors (e.g. Alexander & Alkema, 2022). We therefore conclude that the model adequately depicts the uncertainty of the population forecasts. Furthermore, the MAPE value is relatively small, indicating that the forecasted population closely approximates the true observed population. Overall, these results suggest that our model performs well in forecasting the age-specific population, albeit only validated in the short term. To thoroughly evaluate the performance of long-term age-specific forecasts, additional data are required.

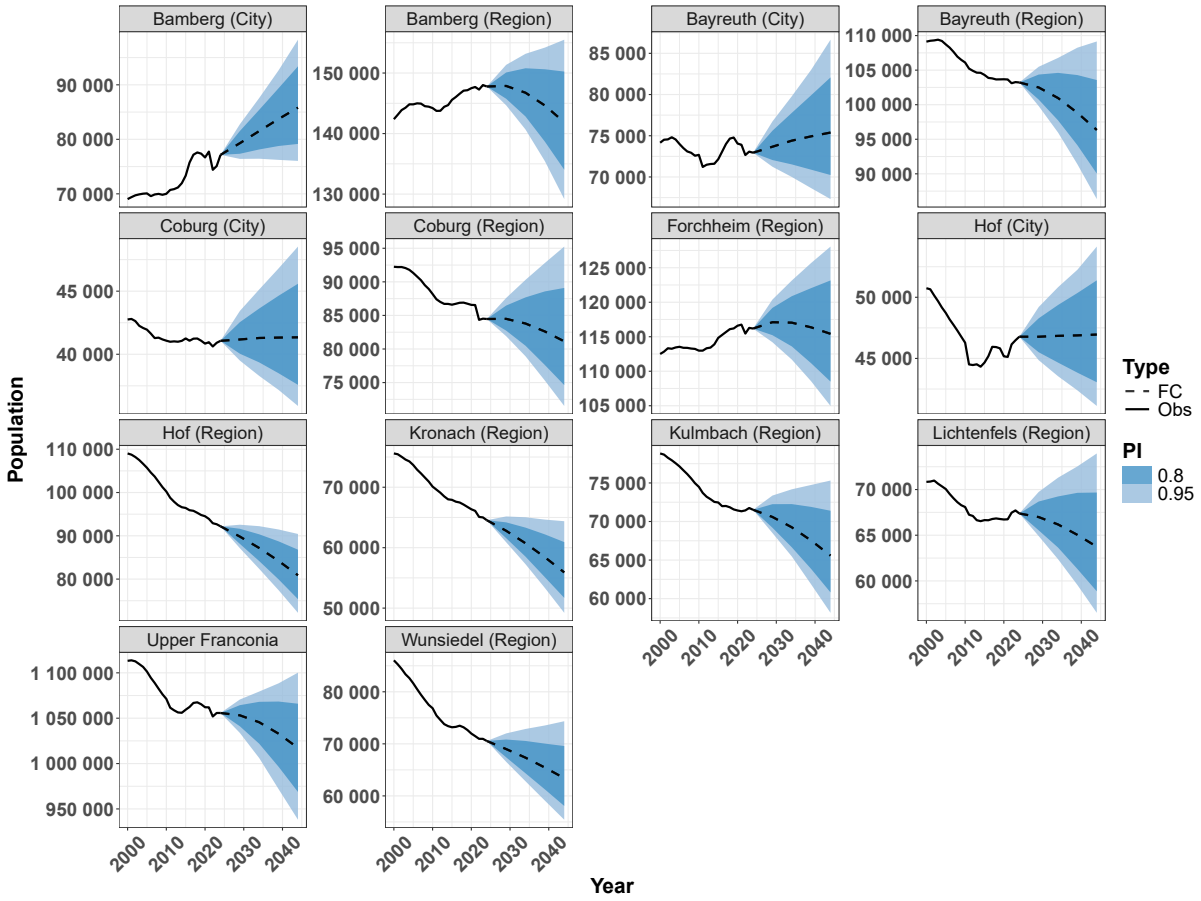
3.7.2 Results

We run our model and obtain population forecasts, disaggregated by age, region, and sex, for the period from 2024 up to 2044 in five-year intervals. The total population projections, including uncertainty intervals for all regions, are presented in Fig. 3.3. In this figure, the black solid line represents the observed values until 2024, while the dashed line indicates the median forecasts, with associated uncertainty intervals depicted in varying shades of blue.

Starting with a total population of 1 055 758 in 2024, the overall number of people in Upper Franconia is expected to slightly decrease by approximately 4% over the projection period. The median forecast of the total population in 2044 is 1 010 867 with an 80% prediction range of 968 694 and 1 065 963. However, this pattern is not uniform across all regions. Forecast results of the province as well as all regions separately are shown in Fig. 3.3.

Generally, we observe three distinct types of patterns that appear to associate with region type: Long-term decline, long-term population growth, and constant population with the possibility of either growth or decline. More precisely, Bavaria, Germany, comprises of two different types of NUTS-3 regions: First, there are regions where multiple smaller cities and villages share an administrative body, denoted as "Region" in Fig. 3.3. Second, there are city regions, denoted as "City" in Fig. 3.3, where a single larger city has its own administrative boundary. We anticipate these city regions to behave differently from their counterparts for two main reasons: First, these city regions can be considered

Figure 3.3: Population forecasts for all regions in Upper Franconia



Note: The observed data is shown by the solid line, while the median forecasts are given by the dashed line. The 80% prediction intervals are plotted in the darker shade of blue while 95% prediction intervals are shown in the lighter shade of blue

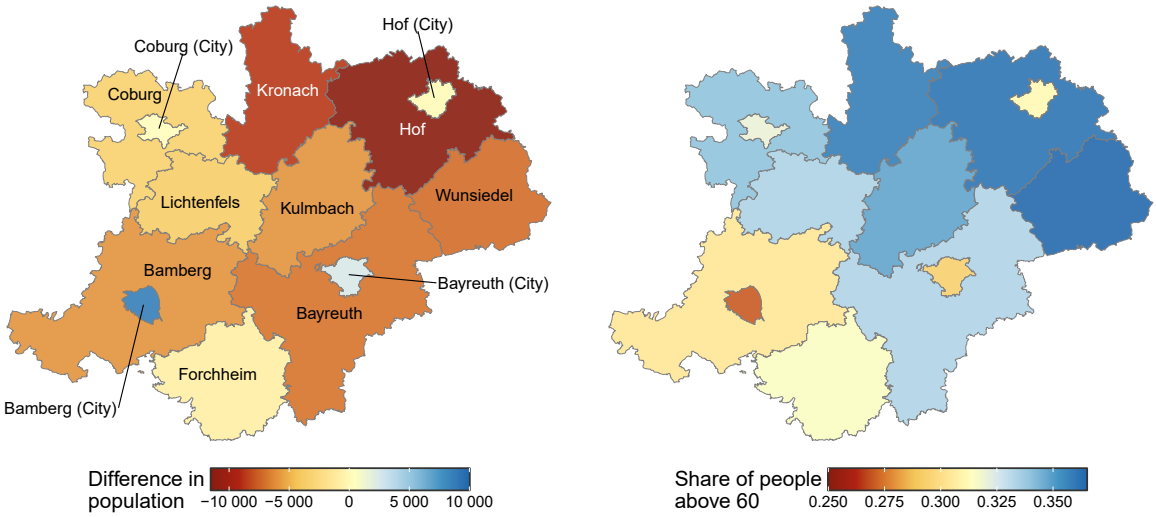
the economic engines of Upper Franconia, as evidenced by their higher GDP values (cf. Fig. C.2). We therefore expect them to have a higher number of net-migration. Second, all of these city regions are considered medium-sized cities (50 000 – 100 000 inhabitants), which have generally experienced population growth in Germany over the last fifteen years, while rural municipalities have seen a decline (Wolff et al., 2022). Examining the population forecasts for these city regions, we observe either likely population growth (Bamberg, Bayreuth) or a constant population (Coburg, Hof), though with a considerable degree of uncertainty.

For the non-city regions, the pattern is quite different. Most are expected to experience a population decline due to low net-migration and a high proportion of older people. In these cases, the low net-migration cannot counteract the negative difference between the expected number of deaths and births, leading to a shrinking population. To better illustrate the negative relationship between population growth and age structure, we plotted the difference in forecasted median population between 2044 and 2024 alongside the share of people over 60 years in 2024 on a map (see Fig. 3.4). This visualization

clearly shows that regions with a high share of older people in 2024 (primarily in the northeast of Upper Franconia) are more likely to experience population decline in the future.

Furthermore, two regions particularly stand out: The region of Bamberg and Forchheim. Both have experienced a fairly constant population growth over the past 20 years, a trend expected to continue in the immediate future. However, this pattern is projected to reverse according to the median forecasts. That is, after a period of growth, a population decline becomes more likely, especially in the region of Bamberg. This is due to an anticipated increase in deaths that is expected to occur in the future. The reason being, that even though these regions have a younger age structure than other members of the "Region" group, people aged between 50–60 still constitute the highest share of the population. In 20 years, however, these individuals will be between 70–80, leading to an increase in expected deaths that projected net-migration may not be able to counteract, thereby reversing the trend of population growth.

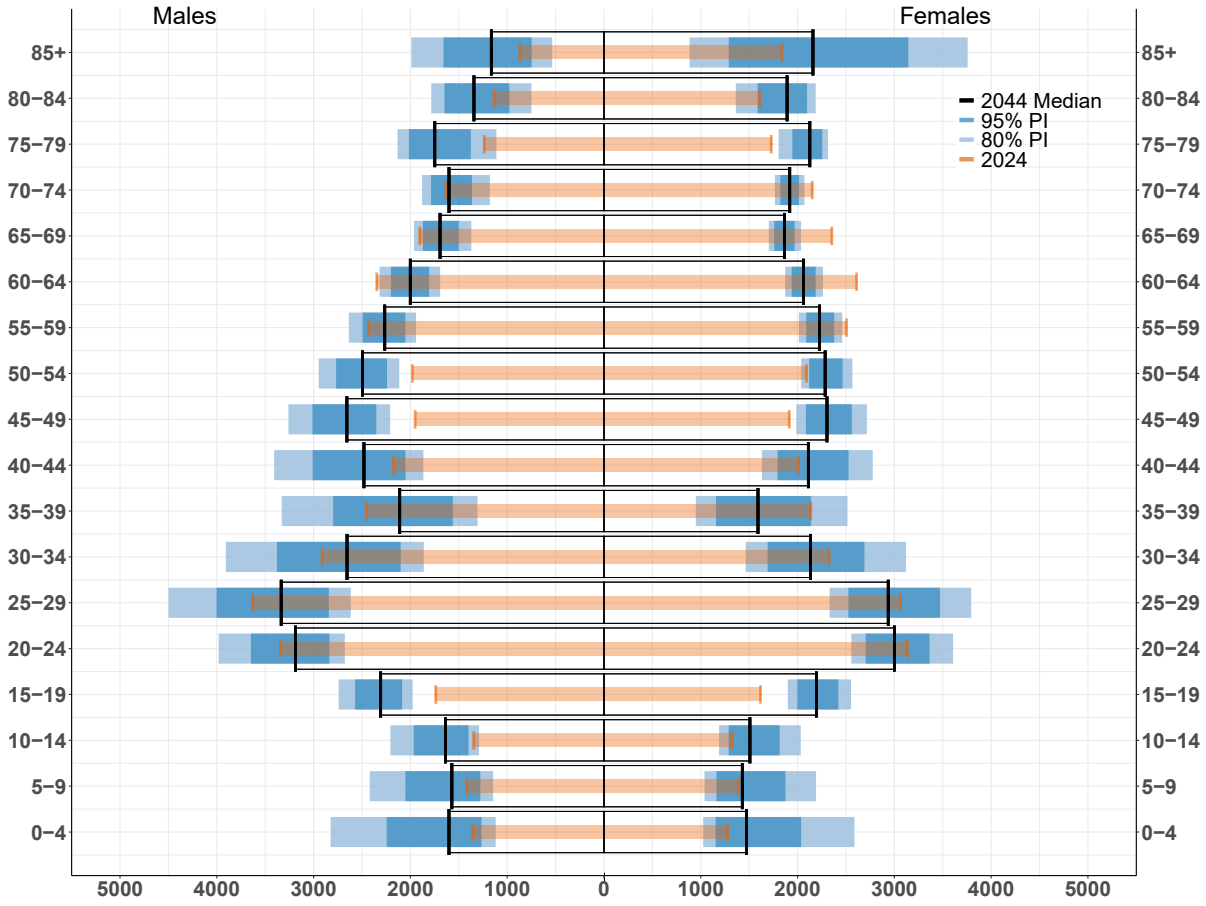
Figure 3.4: Difference in population between 2024 and median forecasts for 2044



Note: The information (Region) was dropped to shorten the name of certain regions for a more compact visualization

In addition to serving as an economic hub, several of the city regions – such as Bamberg, Bayreuth, and Coburg – are home to universities, contributing to a significant population of college-aged residents. It is not expected that this population age-structure will change. Rather, the share of people between 20–30 remains constant over time due to high in-migration counts in early adulthood and high out-migration counts after having graduated university. Looking at the probabilistic population pyramid for the city of Bayreuth (cf. Fig. 3.5), we see that the total amount of people between 20–30 remain relatively constant over time. Thus, our migration model effectively preserves the age-specific shape of these regions into the future.

Figure 3.5: Population pyramid for Bayreuth (City)



3.8 Discussion

In this article, we have developed a probabilistic forecasting method of age-specific population including age-specific mortality, fertility, and net-migration rates for small areas. We introduced new approaches to forecast both regional age-specific fertility rates and regional net-migration counts. Specifically, we extended the standard Lee-Carter model for age-specific fertility rates by including an age-region interaction term, which resulted in improved predictions across both point measures and scoring rules. Additionally, we incorporated the concept of skewed error terms to better capture and predict future migration shocks. Our findings demonstrate that using skewed error terms led to more accurate predictions of net-migration counts compared to the standard approach that relies on symmetric normal errors. Lastly, we introduced a new method to estimate the age-proportion of fertility and the age-specific migration schedule using a Dirichlet regression.

The model was applied to real data to obtain population forecasts for 13 heterogeneous regions in Bavaria, Germany. The long-term predictions (2024–2044) reveal heteroge-

neous regional trends, differentiating between stable or growing city populations driven by economic factors and a high share of students, as well as declining populations in non-city regions attributed to aging and insufficient net-migration. While we did not compare the out-of-sample performance of our population forecasts with that of other methods, we found that our proposed methods for each input variable of the cohort-component method consistently outperformed other standard approaches. This is of significant importance, as improved accuracy in forecasting individual components directly translates to more precise population forecasts overall, due to the underlying fundamental equation.

The method that we have proposed is very general in a sense that it can be applied to forecast age-specific population of other small areas, for example, other NUTS-3 regions. In addition, no adjustments are needed when applying it to predict the age-specific population on a national level. Furthermore, it is unclear whether the indirect approach to forecast age-specific fertility rates performs worse, albeit slightly, than the direct approach due to the simple AR(1) model for regional TFRs, the Dirichlet regression for the age proportions, or a mix of both. Adding a regional correlation structure to model regional TFRs might improve predictions (Fosdick & Raftery, 2014), though more research should be done to see if the Dirichlet regression is suitable in forecasting age proportions or not. Nonetheless, it constitutes an interesting addition to the demographic literature.

There are two valid points of criticism regarding our model. First, it can be considered relatively complex and data-intensive. This could pose an entry barrier for some users in statistical agencies, as it involves significant production costs—particularly in terms of staff time—and requires familiarity with both statistical methods and software. These issues should be considered when deciding between competing methods, not just their predictive performance (cf. Smith et al., 2013, ch. 12). Moreover, some evidence suggests that simpler models can perform just as well as more complex ones when it comes to forecasting (Green & Armstrong, 2015). Simpler extrapolation methods, such as autoregressive integrated moving average (ARIMA) models, which are fit directly on the population data, are less data-intensive and therefore provide a valuable alternative. Nevertheless, most of the data needed by our model should be available, at least for Western countries. Furthermore, when predicting population by age group, one must account for shifts in the age structure, which implies that some variant of the cohort-component method is needed (Smith et al., 2013). In addition, the cohort-component method is directly linked to the fundamental population equation, the cornerstone of demographic theory. This positions the method at the core of population studies, providing a solid theoretical foundation. Methods that use the cohort-component model therefore offer theoretical advantages over other approaches that lack this foundation (Burch, 2018). Second, our proposed method is computationally demanding, as it requires repeated sampling from a complex posterior distribution. However, this enables the generation of probabilistic

population forecasts, including credibility intervals. A quicker and lower-cost alternative based on cohort-change ratios, such as the Hamilton-Perry model (Hamilton & Perry, 1962), is deterministic in nature and produces only point forecasts. Moreover, adopting a frequentist approach for parameter estimation in our proposed model would reduce the computational burden, although it would limit the results to point forecasts as well.To obtain prediction intervals in a frequentist framework, a bootstrap procedure would be required, significantly increasing computational costs and thus negating the advantage of faster computation times.

Using Bayesian methods we have accounted for many sources of uncertainty, however, we have not considered model uncertainty. That is, we have not accounted for model misspecification when predicting each input variable for the cohort-component method. One approach to explicitly incorporate model uncertainty into the prediction problem is the stacking approach by Yao et al., 2018, with implementations in demographic literature by Barigou et al., 2023 and Goes, 2024. Adding model uncertainty may help both in improving the predictive accuracy of our method and in providing an even more realistic description of the underlying uncertainty.

Finally, we have assumed that all inputs of the cohort-component method – namely mortality, fertility, and net-migration – are independent of each other, which may be a problematic assumption. For example, a significant increase in net-migration in a specific region could impact the fertility pattern of that region in the subsequent years, something not considered in our approach. Additionally, the population is forecasted for each sex separately. Allowing for correlation between sexes – similar to Wiśniowski and Raymer, 2025 – or perhaps more importantly between all inputs is an interesting direction of future research.

Acknowledgments

Julius Goes gratefully acknowledges financial support by the Oberfrankenstiftung (grant FP01054). Moreover, Julius Goes thanks Chen-Hao Hsu for valuable discussion on the topic of fertility. Finally, the authors would like to thank the three anonymous reviewers for helpful comments on an earlier version of the manuscript.

C Appendix

C.1 Comparison in-sample fit of migration schedule

We compare the fit of the traditional Rogers Castro curve using the implementation of Yeung et al. (2023) with the Dirichlet model as described in Section 3.5. For the Dirichlet regression having multiple years of migration schedules as input variables helps in reducing the variance of the estimates. The Rogers-Castro model of Yeung et al., 2023, on the other hand, uses a single year as the data input but is proven to work well. To be able to compare both methods, we estimated the migration schedule for each region using the mean values of the age-distribution, $\bar{R}_{x,r}$, as input for the model of Yeung et al. (2023). Our Dirichlet regression model was trained with all available data, that is, multiple years, and therefore has the advantage of having seen more data. Nevertheless, we generated replicate data using the posterior predictive distribution of both models and compared the in-sample fit of each region for both models in terms of MSE and MAE using the age-schedule of all years as the observed values. While the methods are difficult to compare since they are using different input values, we check if the replicated in-sample data of the Dirichlet regression are on a similar level to that of the Rogers-Castro model. Evaluating the results in Table C.1, we notice two things: First, the performance of both models is very satisfactory, and second, the Dirichlet regression performs on par with than the Bayesian implementation of the Rogers-Castro model. Although the Dirichlet regression is estimated with more data, making the comparison difficult.

Table C.1: Comparison of the in-sample performance of the migration schedule for men

Method	**MAE** $\cdot 10^{-2}$	**RMSE** $\cdot 10^{-2}$	**MAPE**
Out-migration			
Rogers Castro	0.64	0.86	28.96
Dirichlet	0.49	0.71	15.52
In-migration			
Rogers Castro	0.66	0.91	29.38
Dirichlet	0.54	0.82	16.42

Note: MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error

For women, we observe similar results. These are given in Table C.2. We conclude that the Dirichlet model is adequate in describing the age-migration patterns.

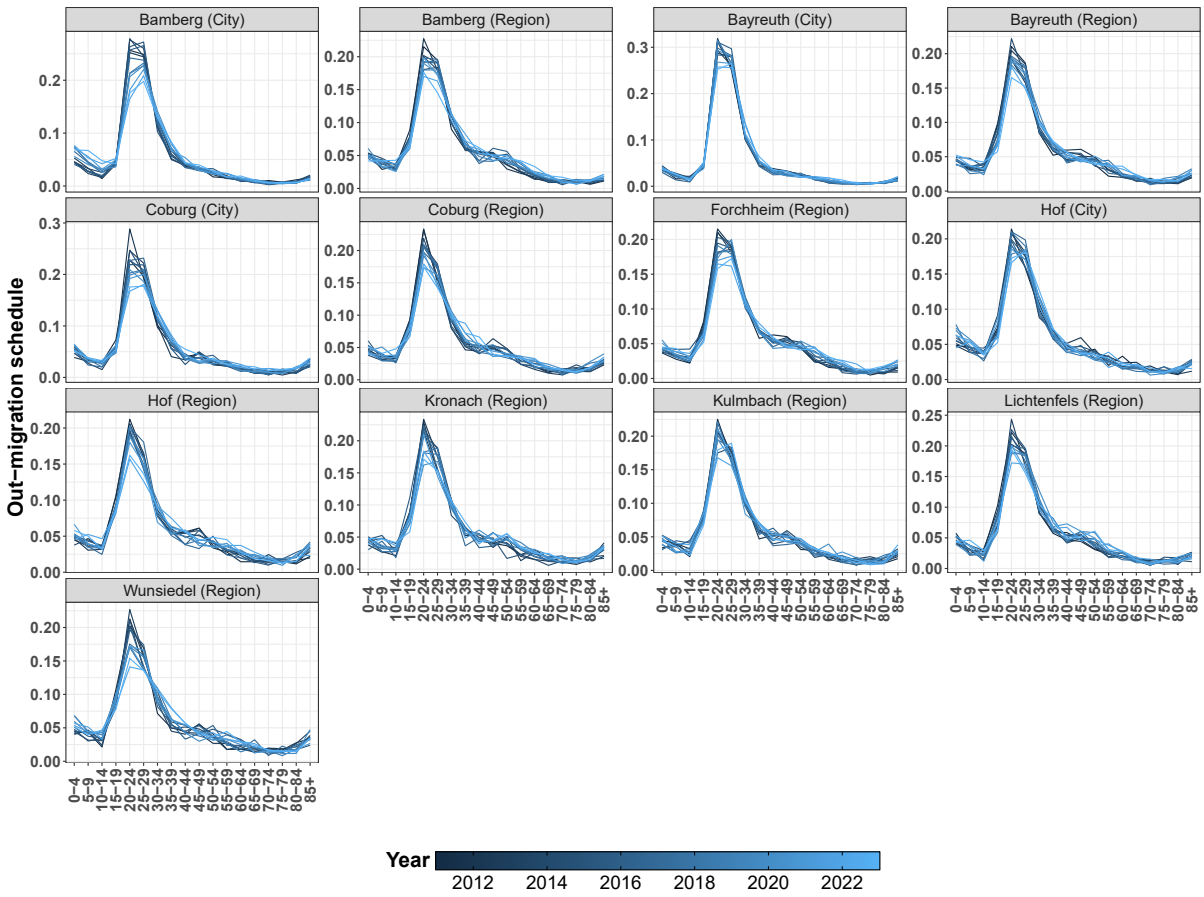
Table C.2: Comparison of the in-sample performance of the migration schedule for women

Method	MAE $\cdot 10^{-2}$	RMSE $\cdot 10^{-2}$	MAPE
Out-migration			
Rogers Castro	0.79	1.02	27.78
Dirichlet	0.56	0.83	14.83
In-migration			
Rogers Castro	0.76	1.04	26.36
Dirichlet	0.59	0.92	15.48

Note: MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error

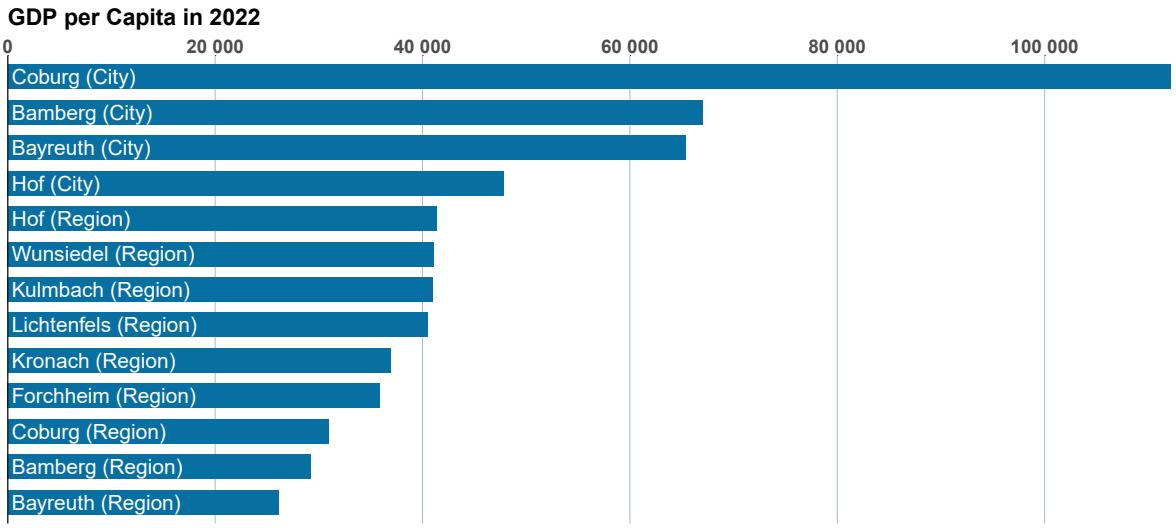
C.2 Graphics of Migration Schedules

Figure C.1: Out-migration schedule over time of women for all regions ins Upper Franconia.



C.3 Additional Figures

Figure C.2: GDP per Capita for all regions in Upper Franconia in 2022



C.4 Additional Tables

Table C.3: Comparison of the out-of-sample performance of the mortality models for males disaggregated by age

AgeGroup	Model	MAE	RMSE	Cov	LogS	RPS
0-4	RH_BYM2	**1.02**	**1.30**	**0.97**	**60.33**	**27.10**
0-4	Lee-Carter	1.03	1.31	1.00	60.51	27.23
5-9	RH_BYM2	**0.24**	**0.39**	**1.00**	**19.75**	**5.94**
5-9	Lee-Carter	0.25	0.40	0.97	21.65	6.16
10-14	RH_BYM2	**0.26**	**0.33**	**1.00**	**14.54**	**3.99**
10-14	Lee-Carter	0.32	0.39	1.00	17.31	4.96
15-19	RH_BYM2	**0.61**	**0.71**	**1.00**	**37.71**	**13.66**
15-19	Lee-Carter	0.74	0.87	1.00	41.83	16.17
20-24	RH_BYM2	**1.08**	**1.42**	**0.92**	**65.31**	**30.84**
20-24	Lee-Carter	1.20	1.46	1.00	67.64	30.90
25-29	Lee-Carter	**1.04**	**1.31**	1.00	**61.64**	**27.65**
25-29	RH_BYM2	1.06	1.38	**0.97**	63.78	29.18
30-34	Lee-Carter	**1.33**	**1.62**	0.97	**69.24**	**34.68**
30-34	RH_BYM2	1.38	1.64	**0.95**	71.62	36.53
35-39	Lee-Carter	**1.12**	**1.37**	1.00	**70.27**	**31.68**
35-39	RH_BYM2	1.21	1.51	**0.95**	76.38	35.16
40-44	RH_BYM2	**1.54**	**2.16**	**0.95**	**80.31**	**43.58**
40-44	Lee-Carter	1.72	2.25	0.92	83.14	46.57
45-49	Lee-Carter	**2.71**	**3.57**	0.97	**101.15**	**73.44**
45-49	RH_BYM2	2.75	3.60	**0.95**	102.75	76.48
50-54	RH_BYM2	**3.04**	**3.74**	**1.00**	**106.10**	**82.31**
50-54	Lee-Carter	3.44	4.27	1.00	112.09	95.07
55-59	RH_BYM2	**4.25**	**5.45**	0.95	**119.19**	**117.37**
55-59	Lee-Carter	4.52	5.64	**1.00**	119.40	118.43
60-64	RH_BYM2	**4.46**	**5.65**	0.92	**124.38**	**130.64**
60-64	Lee-Carter	4.59	5.79	**0.97**	127.42	136.13
65-69	RH_BYM2	**5.09**	**6.35**	**1.00**	**127.70**	**141.77**
65-69	Lee-Carter	7.09	8.54	1.00	140.73	189.48
70-74	RH_BYM2	**5.37**	**6.86**	1.00	**129.06**	**152.67**
70-74	Lee-Carter	8.92	11.32	**0.97**	160.71	227.88
75-79	RH_BYM2	**7.31**	**8.96**	**1.00**	**139.35**	**201.56**
75-79	Lee-Carter	11.75	15.29	1.00	172.27	318.95
80-84	RH_BYM2	**8.30**	**10.58**	**0.95**	**146.91**	**231.47**
80-84	Lee-Carter	9.11	11.51	1.00	150.10	242.82
85+	Lee-Carter	**12.74**	**16.60**	**0.97**	**170.13**	**363.88**
85+	RH_BYM2	13.62	18.00	0.90	175.10	383.49

Note: MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95. LogS = log score. RPS = Ranked probability score

Table C.4: Comparison of the out-of-sample performance of the mortality models for females disaggregated by age

AgeGroup	Model	MAE	RMSE	Cov	LogS	CRPS
0-4	RH_BYM2	**0.85**	**1.09**	**0.97**	**53.49**	**22.44**
0-4	Lee-Carter	0.93	1.13	1.00	55.22	23.78
5-9	RH_BYM2	**0.19**	**0.32**	1.00	**14.25**	**4.16**
5-9	Lee-Carter	0.20	0.34	**0.97**	16.64	4.51
10-14	RH_BYM2	**0.20**	**0.33**	1.00	**14.74**	**4.25**
10-14	Lee-Carter	0.22	0.33	1.00	14.94	4.33
15-19	RH_BYM2	**0.49**	**0.62**	1.00	**30.28**	**10.59**
15-19	Lee-Carter	0.50	0.64	**0.97**	34.35	11.02
20-24	Lee-Carter	**0.46**	**0.49**	1.00	**28.35**	**9.06**
20-24	RH_BYM2	0.48	0.51	1.00	29.33	9.69
25-29	RH_BYM2	**0.70**	**0.87**	**0.97**	**46.54**	**18.09**
25-29	Lee-Carter	0.74	0.92	0.97	49.70	19.58
30-34	RH_BYM2	**0.87**	**1.09**	**0.97**	**50.35**	**21.87**
30-34	Lee-Carter	0.92	1.15	0.97	54.44	23.40
35-39	RH_BYM2	**0.85**	**1.17**	0.97	**55.51**	**23.26**
35-39	Lee-Carter	0.93	1.25	0.97	58.54	25.28
40-44	RH_BYM2	**0.98**	**1.25**	0.97	**62.75**	**26.15**
40-44	Lee-Carter	1.03	1.37	0.97	66.33	28.52
45-49	RH_BYM2	**1.75**	**2.12**	1.00	**82.79**	**46.10**
45-49	Lee-Carter	1.77	2.24	1.00	84.51	47.70
50-54	RH_BYM2	**2.42**	**3.05**	1.00	**98.81**	**68.37**
50-54	Lee-Carter	2.60	3.37	**0.97**	102.19	72.12
55-59	RH_BYM2	**2.91**	**3.70**	**0.92**	**108.23**	**81.98**
55-59	Lee-Carter	3.36	4.25	0.95	114.29	93.84
60-64	RH_BYM2	**3.39**	**4.40**	**0.95**	**111.82**	**94.45**
60-64	Lee-Carter	3.93	5.17	0.97	119.45	111.31
65-69	RH_BYM2	**4.17**	**5.27**	**0.95**	**121.87**	**116.92**
65-69	Lee-Carter	4.59	5.75	0.97	125.91	127.30
70-74	RH_BYM2	**6.37**	**8.69**	0.87	**148.12**	**188.55**
70-74	Lee-Carter	6.79	9.25	**0.90**	153.14	192.52
75-79	RH_BYM2	**6.72**	**8.27**	1.00	**136.60**	**186.23**
75-79	Lee-Carter	11.20	14.68	1.00	179.10	312.43
80-84	RH_BYM2	**9.09**	**11.96**	0.90	**152.66**	**256.78**
80-84	Lee-Carter	11.37	14.63	**0.97**	162.98	315.10
85+	Lee-Carter	**20.94**	**26.51**	**0.97**	**201.22**	**614.16**
85+	RH_BYM2	22.47	27.91	1.00	203.52	629.86

Note: MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95. LogS = log score. RPS = Ranked probability score

Table C.5: Comparison of the out-of-sample performance of the fertility models disaggregated by age

AgeGroup	**Model**	**MAE**	**RMSE**	**MAPE**	**Cov**	**LogS**$\cdot 10^{-2}$	**RPS**$\cdot 10^{-2}$
15-19	Direct Approach	3.88	4.83	39.16	0.90	3.12	2.87
15-19	Indirect Approach	**3.70**	**4.63**	37.00	0.90	**3.06**	**2.77**
15-19	Lee-Carter	3.84	4.79	**36.62**	**0.93**	3.06	2.80
20-24	Direct Approach	9.74	12.62	14.57	**0.94**	4.16	7.43
20-24	Indirect Approach	**9.11**	**11.80**	**13.73**	0.96	**4.10**	**7.02**
20-24	Lee-Carter	11.85	14.75	17.38	0.97	4.32	8.91
25-29	Direct Approach	24.64	30.13	12.12	0.79	5.22	18.78
25-29	Indirect Approach	**23.38**	**29.50**	**11.28**	**0.82**	**5.09**	**17.77**
25-29	Lee-Carter	30.44	38.10	14.58	0.69	5.50	23.55
30-34	Direct Approach	29.46	37.35	11.07	0.85	5.37	22.13
30-34	Indirect Approach	31.43	39.62	11.60	0.85	5.46	23.77
30-34	Lee-Carter	**28.11**	**35.52**	**10.81**	**0.93**	**5.20**	**20.66**
35-39	Direct Approach	**15.67**	**21.29**	**11.57**	**0.93**	**4.59**	**12.15**
35-39	Indirect Approach	16.42	22.40	11.87	0.92	4.60	12.62
35-39	Lee-Carter	25.65	35.43	24.33	0.98	4.79	15.65
40-44	Direct Approach	5.20	6.44	20.29	**0.97**	3.43	3.88
40-44	Indirect Approach	**5.04**	**6.39**	**19.21**	0.96	**3.42**	**3.84**
40-44	Lee-Carter	6.56	9.93	34.18	1.00	3.56	4.23

Note: MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95. LogS = log score. RPS = ranked probability score

Table C.6: Comparison of the out-of-sample performance of age-specific population for Upper Franconia of our model disaggregated by age and sex

Sex	Age group	MAE	RMSE	MAPE	rel-HW	Coverage
female	0-4	103.18	150.75	6.55	12.77	0.85
female	5-9	34.75	46.27	2.12	5.51	1.00
female	10-14	46.03	52.54	2.87	4.97	0.77
female	15-19	92.01	126.44	5.41	5.89	0.46
female	20-24	146.05	197.86	7.41	10.07	0.77
female	25-29	122.17	222.97	5.09	11.16	0.92
female	30-34	105.41	149.56	4.31	8.73	0.92
female	35-39	134.36	172.67	5.35	5.95	0.54
female	40-44	63.46	71.49	2.91	4.62	0.92
female	45-49	35.56	41.33	1.74	4.20	1.00
female	50-54	33.73	41.48	1.34	3.33	0.92
female	55-59	31.99	40.79	0.97	2.35	0.92
female	60-64	36.91	42.46	1.25	1.91	0.85
female	65-69	34.62	49.40	1.16	1.83	0.77
female	70-74	23.46	29.48	0.87	2.06	1.00
female	75-79	33.59	41.11	1.92	4.17	1.00
female	80-84	30.89	37.23	1.78	6.49	1.00
female	85+	52.04	75.14	2.80	15.47	1.00
male	0-4	101.61	161.53	6.32	11.86	0.85
male	5-9	58.77	79.92	3.48	5.44	0.92
male	10-14	38.76	43.50	2.24	4.73	0.92
male	15-19	96.09	120.80	5.59	5.85	0.54
male	20-24	119.63	176.34	5.42	10.92	0.92
male	25-29	203.95	275.83	7.75	13.45	0.77
male	30-34	133.80	182.50	4.88	11.58	0.92
male	35-39	99.08	173.50	3.46	8.36	0.92
male	40-44	58.06	68.32	2.42	6.60	1.00
male	45-49	54.33	68.31	2.41	5.88	0.92
male	50-54	45.36	57.89	1.64	4.64	0.92
male	55-59	56.06	68.09	1.74	3.08	0.77
male	60-64	52.43	62.45	1.55	2.73	0.92
male	65-69	40.33	54.16	1.52	3.36	0.85
male	70-74	32.06	38.07	1.52	5.05	1.00
male	75-79	44.53	50.42	3.31	7.05	1.00
male	80-84	51.83	64.37	4.07	8.37	0.92
male	85+	95.32	115.86	8.87	13.08	0.85

Note: MAE = mean absolute error. RMSE = root mean squared error. MAPE = mean absolute percentage error. Coverage denotes the coverage for a nominal value of 0.95. rel-HW denotes the relative half width and is defined as rel-HW $= 100 \cdot \frac{(y_u - y_l)/2}{E(y)}$, with y_u and y_l denoting the upper respectively lower prediction interval and $E(y)$ the mean forecast

C.5 Cumulative distribution and quantile function of skewed distribution

Consider a univariate random variable with probability density function $f(\cdot)$, that is continuous, unimodal and symmetric around 0, for example, a centered normal or student-t distribution. A parameter $\gamma \in (0, \infty)$ accounts for the degree of skewness which generates

the following class of skewed distributions:

$$p(x|\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \left[f\left(\frac{x}{\gamma}\right) \mathbb{1}_{[0,\infty)}(x) + f\left(x\gamma\right) \mathbb{1}_{(-\infty,0)}(x) \right].$$

Here, $\mathbb{1}_{[0,\infty)}(\varepsilon)$ denotes an indicator function which equals one if ε lies in the interval $[0, \infty)$ and zero otherwise.

The cumulative distribution function, CDF, which we will denote by $P(x|\gamma)$, can be calculated using integration by substitution. Recall that integration by substitution is given by

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(t)dt.$$

Then to obtain the CDF, we have

$$P(x|\gamma) = \frac{2}{\gamma + 1/\gamma} \cdot \left[\int_{-\infty}^{\min(0,x)} f(u\gamma)du + \int_0^x f\left(\frac{u}{\gamma}\right) \mathbb{1}_{[0,\infty)}(x)du \right].$$

We start by looking at $x < 0$ which results in

$$\begin{aligned} P(x|\gamma) &= \frac{2}{\gamma + 1/\gamma} \cdot \int_{-\infty}^{x} f(u\gamma)du \\ &= \frac{2}{\gamma + 1/\gamma} \cdot \int_{-\infty}^{x\gamma} \frac{f(t)}{\gamma} dt \\ &= \frac{2}{\gamma^2 + 1} \cdot F(x\gamma), \end{aligned}$$

where $F(x)$ denotes the CDF of $f(x)$. Next, for $x \geq 0$ we have

$$\begin{aligned} P(x|\gamma) &= P(0|\gamma) + \frac{2}{\gamma + 1/\gamma} \cdot \int_0^x f\left(\frac{u}{\gamma}\right) du \\ &= P(0|\gamma) + \frac{2}{\gamma + 1/\gamma} \int_0^{\frac{x}{\gamma}} f(t) \cdot \gamma dt \\ &= P(0|\gamma) + \frac{2}{\gamma + 1/\gamma} \cdot \left[\gamma \cdot F\left(\frac{x}{\gamma}\right) - \gamma \cdot F(0) \right]. \end{aligned}$$

Thus,

$$P(x|\gamma) = \begin{cases} \frac{2}{\gamma^2+1} \cdot F(x\gamma) & \text{for } x < 0 \\ \underbrace{\frac{2}{\gamma^2 + 1} \cdot F(0)}_{P(0|\gamma)} + \frac{2\gamma}{\gamma+1/\gamma} \cdot \left[F\left(\frac{x}{\gamma}\right) - 1/2 \right] & \text{for } x \geq 0. \end{cases}$$

The quantile function $G(z|\gamma) = P^{-1}(z|\gamma)$ can be obtained by solving $z = P(x|\gamma)$ for x,

with $P(\cdot|\gamma)$ strictly increasing.

Starting with $x < 0$, we get

$$\begin{aligned} & z = \frac{2}{\gamma^2+1} \cdot F(x\gamma) \\ \Leftrightarrow \quad & \frac{z \cdot (\gamma^2+1)}{2} = F(x\gamma) \\ \Leftrightarrow \quad & x = \frac{1}{\gamma} \cdot F^{-1}\left(\frac{z \cdot (\gamma^2+1)}{2}\right). \end{aligned}$$

Here, $F^{-1}(z)$ denotes the quantile function of $f(.)$. Now we look at $x \geq 0$

$$\begin{aligned} & z = P(0|\gamma) + \frac{2\gamma^2}{\gamma^2+1} \cdot [F(x/\gamma) - 1/2] \\ \Leftrightarrow \quad & \frac{\gamma^2 + 1 \cdot [z - P(0|\gamma)]}{2\gamma^2} = F(x/\gamma) - 1/2 \\ \Leftrightarrow \quad & \frac{\gamma^2 \cdot [(1 + 1/\gamma^2) \cdot (z - P(0|\gamma))]}{\gamma^2 \cdot 2} = F(x/\gamma) - 1/2 \\ \Leftrightarrow \quad & \frac{(1 + 1/\gamma^2) \cdot [z - P(0|\gamma)]}{2} + 1/2 = F(x/\gamma) \\ \Leftrightarrow \quad & x = F^{-1}\left[\frac{\left(1 + \frac{1}{\gamma^2}\right) \cdot [z - P(0|\gamma)]}{2} + 1/2\right] \cdot \gamma \\ \Leftrightarrow \quad & x = F^{-1}\left[\frac{\left(1 + \frac{1}{\gamma^2}\right) \cdot \left[z - \frac{2}{2\gamma^2+2}\right]}{2} + 1/2\right] \cdot \gamma. \end{aligned}$$

Thus, the quantile function $G(z|\gamma)$ is given by

$$G(z|\gamma) = \begin{cases} \frac{1}{\gamma} \cdot F^{-1}\left(\frac{z \cdot (\gamma^2+1)}{2}\right) & \text{for } z < P(0|\gamma) \\ F^{-1}\left[\frac{\left(1+\frac{1}{\gamma^2}\right) \cdot \left[z - \frac{2}{2\gamma^2+2}\right]}{2} + 1/2\right] & \text{for } z \geq P(0|\gamma). \end{cases}$$

C.6 Overview on choice of priors

For the mortality model, we use the following choice of priors

$$
\begin{aligned}
\alpha_x|\alpha_{x-1},\alpha_{x-2};\sigma_\alpha^2 &\sim \mathcal{N}(2\alpha_{x-1}-\alpha_{x-2},\sigma_\alpha^2) \quad \text{and } \alpha_j \sim \mathcal{N}(0,\sigma_\alpha^2) \text{ for } j=1,2\\
\sigma_\alpha &\sim t_5^+(0,1)\\
\left(\beta_1^{(1)},\ldots,\beta_X^{(1)}\right) &\sim \text{Dirichlet}(1,\ldots,1)\\
\left(\beta_1^{(2)},\ldots,\beta_X^{(2)}\right) &\sim \text{Dirichlet}(1,\ldots,1)\\
\kappa_t|\kappa_{t-1},d;\sigma_\kappa^2 &\sim \mathcal{N}(\kappa_{t-1}+d,\sigma_\kappa^2), \quad \text{and } k_1 \sim \mathcal{N}(d,\sigma_k^2)\\
d &\sim \mathcal{N}(0,2^2)\\
\sigma_\kappa &\sim t_5^+(0,1)\\
\gamma_k &\overset{iid}{\sim} \mathcal{N}(0,\sigma_\gamma)\\
\sigma_\gamma &\sim t_5^+(0,1)\\
\nu_r &= \sigma_\nu\left(\sqrt{1-\rho}v_r^* + \sqrt{\rho}u_r^*\right)\\
u_r|u_{-r};\sigma_u^2 &\sim \mathcal{N}\left(\frac{\sum_{r\neq j} w_{rj}u_j}{\sum_{r\neq j} w_{rj}}, \frac{\sigma_u^2}{\sum_{r\neq j} w_{rj}}\right)\\
\sigma_\nu &\sim t_5^+(0,1)\\
\rho &\sim \text{Beta}(0.5,0.5),
\end{aligned}
$$

where ν_r follows a BYM2 model. Here, σ_ν denotes its standard deviation, $\rho \in [0,1]$ defines a mixing parameter how much of the effect is spatially unstructured, given by $v_r^* \sim \mathcal{N}(0,1)$ and how much is due to a scaled spatially structured effect u_r^* which follows a scaled conditional autoregressive model, with symmetric adjacency matrix $\mathbf{W} \in \mathbb{R}^{R\times R}$ with entries w_{rj} denoting one if two regions are neighbors and zero otherwise (see Riebler et al., 2016 for details). In addition $t_5^+(0,1)$ denotes a location scale t-distribution truncated to the interval $[0,\infty)$ with five degrees of freedom, location of zero and scale of one. These are the same priors as used in Goes (2024).

For the direct fertility model, we use the following priors

$$\begin{aligned}
\alpha_x &\overset{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2) \\
\sigma_\alpha &\sim t_5^+(0, 1) \\
\beta_x &\overset{iid}{\sim} \mathcal{N}(0, 5^2) \\
\kappa_t|\kappa_{t-1}, d; \sigma_\kappa^2 &\sim \mathcal{N}(\kappa_{t-1} + d, \sigma_\kappa^2) \quad \text{and } k_1 \sim \mathcal{N}(d, \sigma_k^2) \\
d &\sim \mathcal{N}(0, 2^2) \\
\sigma_\kappa &\sim t_5^+(0, 1) \\
\delta_{x,r} &\overset{iid}{\sim} \mathcal{N}(0, 10) \\
\sigma_\varepsilon &\sim t_5^+(0, 1).
\end{aligned}$$

To model net-migration counts, the following choice of priors were used

$$\begin{aligned}
\mu_r &\overset{iid}{\sim} \mathcal{N}(m_1, s_1) \\
m_1 &\sim \mathcal{N}(0, 200) \\
s_1 &\sim \mathcal{N}^+(0, 100) \\
\varepsilon_{t,r} &\overset{iid}{\sim} \text{Skew-Normal}(\gamma) \\
\gamma &\sim \mathcal{N}(1, 1) \\
\sigma_r &\overset{iid}{\sim} \mathcal{N}^+(m_2, s_2) \\
m_2 &\sim \mathcal{N}(0, 50) \\
s_2 &\sim \mathcal{N}^+(0, 20) \\
a_r &\overset{iid}{\sim} \mathcal{N}(0, 2).
\end{aligned}$$

Bibliography

Alexander, M., & Alkema, L. (2022). A Bayesian cohort component projection model to estimate women of reproductive age at the subnational level in data-sparse settings. *Demography*, *59*(5), 1713–1737. https://doi.org/10.1215/00703370-10216406

Alexander, M., Zagheni, E., & Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, *54*(6), 2025–2041. https://doi.org/10.1007/s13524-017-0618-7

Alexopoulos, A., Dellaportas, P., & Forster, J. J. (2019). Bayesian forecasting of mortality rates by using latent Gaussian models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(2), 689–711. https://doi.org/10.1111/rssa.12422

Alho, J. M., & Spencer, B. D. (2005). *Statistical demography and forecasting*. Springer Science + Business Media.

Antonio, K., Bardoutsos, A., & Ouburg, W. (2015). Bayesian Poisson log-bilinear models for mortality projections with multiple populations. *European Actuarial Journal*, *5*(2), 245–281. https://doi.org/10.1007/s13385-015-0115-6

Azose, J. J., & Raftery, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography*, *52*(5), 1627–1650. https://doi.org/10.1007/s13524-015-0415-0

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data* (2nd ed.). CRC Press.

Barigou, K., Goffard, P.-O., Loisel, S., & Salhi, Y. (2023). Bayesian model averaging for mortality forecasting using leave-future-out validation. *International Journal of Forecasting*, *39*(2), 674–690. https://doi.org/10.1016/j.ijforecast.2022.01.011

Bartolucci, F., Pennoni, F., & Mira, A. (2021). A multivariate statistical approach to predict COVID-19 count data with epidemiological interpretation and uncertainty quantification. *Statistics in Medicine*, *40*(24), 5351–5372. https://doi.org/10.1002/sim.9129

Bayerisches Landesamt für Statistik. (2022a). *Bruttoinlandsprodukt und Bruttowertschöpfung in Bayern 2012 bis 2020* (Statistical Report No. No. 08.2022). Fürth, Germany. https://www.statistischebibliothek.de/mir/receive/BYHeft_mods_00013132

Bayerisches Landesamt für Statistik. (2022b). Dataset 12613-108s. https://www.statistikdaten.bayern.de/genesis/online

Bayerisches Landesamt für Statistik. (2024a). Dataset 12612-005. https://www.statistikdaten.bayern.de/genesis/online

Bayerisches Landesamt für Statistik. (2024b). Dataset 12711-104. https://www.statistikdaten.bayern.de/genesis/online

Bayerisches Landesamt für Statistik. (2025). Dataset 12411-007s. https://www.statistikdaten.bayern.de/genesis/online

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20. https://doi.org/10.1007/BF00116466

Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, *70*(1), 1–19. https://doi.org/10.1080/00324728.2015.1122826

Billari, F. C., Graziani, R., & Melilli, E. (2014). Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm. *Demography*, *51*(5), 1933–1954. https://doi.org/10.1007/s13524-014-0318-5

Bohk-Ewald, C., Li, P., & Myrskylä, M. (2018). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences*, *115*(37), 9187–9192. https://doi.org/10.1073/pnas.1722364115

Bonnet, F., Grigoriev, P., Sauerberg, M., Alliger, I., Mühlichen, M., & Camarda, C.-G. (2024). Spatial disparities in the mortality burden of the COVID-19 pandemic across 569 European regions (2020-2021). *Nature Communications*, *15*(1), 4246. https://doi.org/10.1038/s41467-024-48689-0

Booth, H., & Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, *3*(1-2), 3–43. https://doi.org/10.1017/S1748499500000440

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, *22*(3), 547–581. https://doi.org/10.1016/j.ijforecast.2006.04.001

Bryant, J., & Zhang, J. L. (2016). Bayesian forecasting of demographic rates for small areas: Emigration rates by age, sex, and region in New Zealand, 2014-2038. *Statistica Sinica*, *26*, 1337–1363. https://doi.org/10.5705/ss.2014.200t

Burch, T. K. (2018). *Model-Based Demography*. Springer International Publishing. https://doi.org/10.1007/978-3-319-65433-1

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, *90*(14), 2499–2523. https://doi.org/10.1080/00949655.2020.1783262

Cairns, A. J. G., Blake, D., & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, *73*(4), 687–718. https://doi.org/10.1111/j.1539-6975.2006.00195.x

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., & Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA*, *41*(1), 29–59. https://doi.org/10.2143/AST.41.1.2084385

Cairns, A. J. G., Blake, D. P., Kessler, A., & Kessler, M. (2020). The impact of COVID-19 on future higher-age mortality. *Preprint*. Retrieved September 26, 2023, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3606988

Cameron, M. P., & Poot, J. (2011). Lessons from stochastic small-area population projections: The case of Waikato subregions in New Zealand. *Journal of Population Research*, *28*, 245–265. https://doi.org/10.1007/s12546-011-9056-3

Chen, F.-Y., Yang, S. S., & Huang, H.-C. (2022). Modeling pandemic mortality risk and its application to mortality-linked security pricing. *Insurance: Mathematics and Economics*, *106*, 341–363. https://doi.org/10.1016/j.insmatheco.2022.06.002

Chen, H., & Cox, S. H. (2009). Modeling mortality with jumps: Applications to mortality securitization. *Journal of Risk and Insurance*, *76*(3), 727–751. https://doi.org/10.1111/j.1539-6975.2009.01313.x

Congdon, P. (2014). Estimating life expectancies for U.S. small areas: A regression framework. *Journal of Geographical Systems*, *16*(1), 1–18. https://doi.org/10.1007/s10109-013-0177-4

Cox, S. H., Lin, Y., & Wang, S. (2006). Multivariate exponential tilting and pricing implications for mortality securitization. *Journal of Risk and Insurance*, *73*(4), 719–736. https://doi.org/10.1111/j.1539-6975.2006.00196.x

Czado, C., Delwarde, A., & Denuit, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, *36*(3), 260–284. https://doi.org/10.1016/j.insmatheco.2005.01.001

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model sssessment for count data. *Biometrics*, *65*(4), 1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Cortes, C. W., Rodríguez, A., Lang, D. T., Zhang, W., Paganin, S., Hug, J., Zhang, W., Babu, J., Ponisio, L., & Sujan, P. (2024). nimble: MCMC, particle

filtering, and programmable hierarchical modeling [Version 1.2.1]. https://cran.r-project.org/web/packages/nimble/index.html

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413.

Dharamshi, A., Alexander, M., Winant, C., & Barbieri, M. (2025). Jointly estimating subnational mortality for multiple populations. *Demographic Research*, *52*, 71–110. https://doi.org/10.4054/DemRes.2025.52.3

Enchev, V., Torsten Kleinow, & Cairns, A. J. G. (2017). Multi-population mortality models: Fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, *2017*(4), 319–342. https://doi.org/10.1080/03461238.2015.1133450

Ezzati, M., Friedman, A. B., Kulkarni, S. C., & Murray, C. J. L. (2008). The reversal of fortunes: Trends in county mortality and cross-county mortality disparities in the United States. *PLoS Medicine*, *5*(4), e66. https://doi.org/10.1371/journal.pmed.0050066

Faust, J. S., Du, C., Renton, B., Liang, C., Chen, A. J., Li, S.-X., Lin, Z., Nunez-Smith, M., & Krumholz, H. M. (2022). Two years of COVID-19: Excess mortality by age, region, gender, and race/ethnicity in the United States during the COVID-19 pandemic, March 1, 2020, through February 28, 2022. https://doi.org/10.1101/2022.08.16.22278800

Federal Ministry of Interior and Community & Federal Office for Migration and Refugees. (2024). *Migration peport of the federal government 2022. Executive summary* (tech. rep.). BAMF. Berlin, Nueremberg. https://doi.org/10.48570/bamf.fz.kurz.mb2022.en.2024.migrationreport.1.0

Federal Office for Migration and Refugees. (2016). *Migration report 2015. Central conclusions* (tech. rep.). BAMF. Nueremberg. https://www.bamf.de/SharedDocs/Anlagen/EN/Forschung/Migrationsberichte/migrationsbericht-2015-zentrale-ergebnisse.html?nn=447186

Federal Office for Migration and Refugees. (2021). *Evaluation of AnkER facilities and functionally equivalent facilities* (Research Report No. 37). BAMF. Nueremberg. https://www.bamf.de/SharedDocs/Anlagen/EN/Forschung/Forschungsberichte/fb37-evaluation-anker-fg-einrichtungen.html

Felice, E., Pujol Andreu, J., & D'Ippoliti, C. (2016). GDP and life expectancy in Italy and Spain over the long run: A time-series approach. *Demographic Research*, *35*, 813–866. https://doi.org/10.4054/DemRes.2016.35.28

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Van Elsland, S., ... Ghani, A. (2020). *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand* (tech. rep. No. 9). Imperial College London. https://doi.org/10.25561/77482

Fernández, C., & Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, *93*(441), 359–371. https://doi.org/10.1080/01621459.1998.10474117

Fosdick, B., & Raftery, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demographic Research*, *30*, 1011–1034. https://doi.org/10.4054/DemRes.2014.30.35

Gagnon, A., Miller, M. S., Hallman, S. A., Bourbeau, R., Herring, D. A., Earn, D. J., & Madrenas, J. (2013). Age-specific mortality during the 1918 Influenza pandemic: Unravelling the mystery of high young adult mortality. *PLoS ONE*, *8*(8), e69586. https://doi.org/10.1371/journal.pone.0069586

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*(3), 115–146. https://doi.org/10.2307/2986645

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

Goes, J. (2024). Bayesian forecasting of mortality rates for small areas using spatiotemporal models. *Demography*, *61*(2), 439–462. https://doi.org/10.1215/00703370-11212716

Goes, J., Barigou, K., & Leucht, A. (2025). Bayesian mortality modelling with pandemics: A vanishing jump approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *74*(4), 1150–1182. https://doi.org/10.1093/jrsssc/qlaf018

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678–1685. https://doi.org/10.1016/j.jbusres.2015.03.026

Gueorguieva, R., Rosenheck, R., & Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics and Data Analysis*, *52*(12), 5344–5355. https://doi.org/10.1016/j.csda.2008.05.030

Haberman, S., & Renshaw, A. (2012). Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and Economics*, *50*(3), 309–333. https://doi.org/10.1016/j.insmatheco.2011.11.005

Hamilton, C. H., & Perry, J. (1962). A Short Method for projecting population by age from one decennial census to another. *Social Forces*, *41*(2), 163–170. https://doi.org/10.2307/2573607

Heuer, C. (1997). Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, *53*, 161–177. https://doi.org/10.2307/2533105

Hobcraft, J., Menken, J., & Preston, S. (1982). Age, period, and cohort effects in demography: A review. *Population Index*, *48*, 4–43. https://doi.org/10.2307/2736356

Homburg, A., Weiß, C. H., Alwan, L. C., Frahm, G., & Göb, R. (2021). A performance analysis of prediction intervals for count time series. *Journal of Forecasting*, *40*(4), 603–625. https://doi.org/10.1002/for.2729

Hunt, A., & Blake, D. (2020). Identifiability in age/period mortality models. *Annals of Actuarial Science*, *14*(2), 461–499. https://doi.org/10.1017/S1748499520000111

Jentsch, C., & Leucht, A. (2016). Bootstrapping sample quantiles of discrete data. *Annals of the Institute of Statistical Mathematics*, *68*(3), 491–539. https://doi.org/10.1007/s10463-015-0503-3

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, *90*(12), 1–37. https://doi.org/10.18637/jss.v090.i12

Keilman, N. (2019). Erroneous population forecasts. In T. Bengtsson & N. Keilman (Eds.), *Old and New Perspectives on Mortality Forecasting* (pp. 95–111). Springer. https://doi.org/10.1007/978-3-030-05075-7_9

Keilman, N. (2020). Evaluating Probabilistic Population Forecasts. *Economie et Statistique / Economics and Statistics*, (520-521).

Keyfitz, N. (1972). On future population. *Journal of the American Statistical Association*, *67*(338), 347–363. https://doi.org/10.1080/01621459.1972.10482386

Krüger, F., Lerch, S., Thorarinsdottir, T., & Gneiting, T. (2021). Predictive inference based on Markov chain Monte Carlo output. *International Statistical Review*, *89*(2), 274–301. https://doi.org/10.1111/insr.12405

Kuang, D., Nielsen, B., & Nielsen, J. P. (2008). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika*, *95*(4), 987–991.

Lee, R. D. (1993). Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level. *International Journal of Forecasting*, *9*(2), 187–202. https://doi.org/10.1016/0169-2070(93)90004-7

Lee, R. D. (1998). Probabilistic approaches to population forecasting. In W. Lutz, J. W. Vaupel, & D. A. Ahlburg (Eds.), *Frontiers of Population Forecasting. A supplement to Population and Development Review* (pp. 156–190, Vol. 24). Population

Council, Wiley. Retrieved July 3, 2025, from https://www.jstor.org/stable/28080 55

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/10.2307 /2290201

Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, *33*(3), 183–212. https://doi.org/10.2307/2332297

Li, J. S.-H., & Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, *15*(2), 177–200. https://doi.org/10.1080/10920277 .2011.10597616

Li, J. S.-H., Zhou, K. Q., Zhu, X., Chan, W.-S., & Chan, F. W.-H. (2019). A Bayesian approach to developing a stochastic mortality model for China. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(4), 1523–1560. https ://doi.org/10.1111/rssa.12473

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.or g/10.1353/dem.2005.0021

Liu, Y., & Li, J. S.-H. (2015). The age pattern of transitory mortality jumps and its impact on the pricing of catastrophic mortality bonds. *Insurance: Mathematics and Economics*, *64*, 135–150. https://doi.org/10.1016/j.insmatheco.2015.05.005

Lorentzen, P., McMillan, J., & Wacziarg, R. (2008). Death and development. *Journal of Economic Growth*, *13*(2), 81–124. https://doi.org/10.1007/s10887-008-9029-3

Luy, M., Di Giulio, P., Di Lego, V., Lazarevič, P., & Sauerberg, M. (2020). Life expectancy: Frequently used, but hardly understood. *Gerontology*, *66*(1), 95–104. https://doi.org/10.1159/000500955

Ma, Y., Genton, M. G., & Parzen, E. (2011). Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, *63*(2), 227–243. https://doi.org/10.1007/s10463-008-0215-z

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., & Clark, S. (2015). Space–time smoothing of complex survey data: Small area estimation for child mortality. *The Annals of Applied Statistics*, *9*(4), 1889–1905. https://doi.or g/10.1214/15-AOAS872

Mitchell, D., Brockett, P., Mendoza-Arriaga, R., & Muthuraman, K. (2013). Modeling and forecasting mortality rates. *Insurance: Mathematics and Economics*, *52*(2), 275–285. https://doi.org/10.1016/j.insmatheco.2013.01.002

Monod, M., Blenkinsop, A., Xi, X., Hebert, D., Bershan, S., Tietze, S., Baguelin, M., Bradley, V. C., Chen, Y., Coupland, H., Filippi, S., Ish-Horowicz, J., McManus,

M., Mellan, T., Gandy, A., Hutchinson, M., Unwin, H. J. T., van Elsland, S. L., Vollmer, M. A. C., … Imperial College COVID-19 Response Team. (2021). Age groups that sustain resurging COVID-19 epidemics in the United States. *Science (New York, N.Y.)*, *371*(6536), eabe8372. https://doi.org/10.1126/science.abe8372

Murray, C. J. L., Kulkarni, S. C., Michaud, C., Tomijima, N., Bulzacchelli, M. T., Iandiorio, T. J., & Ezzati, M. (2006). Eight Americas: Investigating mortality disparities across races, counties, and race-counties in the United States. *PLoS Medicine*, *3*(9), e260. https://doi.org/10.1371/journal.pmed.0030260

Myrskylä, M., Goldstein, J. R., & Cheng, Y.-h. A. (2013). New cohort fertility forecasts for the developed world: Rises, falls, and reversals. *Population and Development Review*, *39*(1), 31–56. https://doi.org/10.1111/j.1728-4457.2013.00572.x

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113–162). Boca Raton, FL: CRC Press.

Neelon, B., Ghosh, P., & Loebs, P. F. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *176*(2), 389–413. https://doi.org/10.1111/j.1467-985X.2012.01039.x

Ng, K. W., Tian, G.-L., & Tang, M.-L. (2011). *Dirichlet and related distributions: Theory, methods and applications*. Wiley.

Ocaña-Riola, R., & Mayoral-Cortés, J. M. (2010). Spatio-temporal trends of mortality in small areas of southern Spain. *BMC Public Health*, *10*(1), 26. https://doi.org/10.1186/1471-2458-10-26

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A. T., Azman, A. S., Paireau, J., Fontanet, A., Cauchemez, S., & Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, *590*(7844), 140–145. https://doi.org/10.1038/s41586-020-2918-0

Pedroza, C. (2006). A Bayesian forecasting model: Predicting U.S. male mortality. *Biostatistics*, *7*(4), 530–550. https://doi.org/10.1093/biostatistics/kxj024

Pitacco, E. (2009). *Modelling longevity dynamics for pensions and annuity business*. Oxford University Press.

Pizzato, M., Gerli, A. G., Vecchia, C. L., & Alicandro, G. (2024). Impact of COVID-19 on total excess mortality and geographic disparities in Europe, 2020–2023: A spatio-temporal analysis. *The Lancet Regional Health – Europe*, *44*. https://doi.org/10.1016/j.lanepe.2024.100996

Preston, S. H., Heuveline, P., & Guillot, M. (2000). *Demography: Measuring and modeling population processes*. Blackwell.

Raftery, A. E., & Ševčíková, H. (2023). Probabilistic population forecasting: Short to very long-term. *International Journal of Forecasting*, *39*(1), 73–97. https://doi.org/10.1016/j.ijforecast.2021.09.001

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd). John Wiley & Sons. https://doi.org/10.1002/9781118735855

Rau, R., & Schmertmann, C. P. (2020). District-Level life expectancy in Germany. *Deutsches Ärzteblatt international*. https://doi.org/10.3238/arztebl.2020.0493

Renshaw, A., & 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

Richards, S. J. (2024). Robust mortality forecasting in the presence of outliers. *British Actuarial Journal*, *29*, 1–23. https://doi.org/10.1017/S1357321724000175

Richardson, K., Jatrana, S., Tobias, M., & Blakely, T. (2013). Migration and pacific mortality: Estimating migration effects on pacific mortality rates using Bayesian models. *Demography*, *50*(6), 2053–2073. https://doi.org/10.1007/s13524-013-0234-0

Riebler, A., & Held, L. (2017). Projecting the future burden of cancer: Bayesian age–period–cohort analysis with integrated nested Laplace approximations. *Biometrical Journal*, *59*(3), 531–549. https://doi.org/10.1002/bimj.201500263

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, *25*(4), 1145–1165. https://doi.org/10.1177/0962280216660421

Robben, J., & Antonio, K. (2024). Catastrophe risk in a stochastic multi-population mortality model. *Journal of Risk and Insurance*. https://doi.org/10.1111/jori.12470

Rogers, A., & Castro, L. J. (1981). *Model migration schedules* (tech. rep. No. RR-81-030). International Institute for Applied Systems Analysis. Laxenberg, Austria. Retrieved December 18, 2023, from https://iiasa.dev.local/

Rogers, A. (1990). Requiem for the net migrant. *Geographical Analysis*, *22*(4), 283–300. https://doi.org/10.1111/j.1538-4632.1990.tb00212.x

Schmertmann, C. P., Cavenaghi, S. M., Assunção, R. M., & and Potter, J. E. (2013). Bayes plus Brass: Estimating total fertility for many small areas from sparse census data. *Population Studies*, *67*(3), 255–273. https://doi.org/10.1080/00324728.2013.795602

Schmertmann, C. P., & Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, *55*(4), 1363–1388. https://doi.org/10.1007/s13524-018-0695-2

Schnürch, S., Kleinow, T., Korn, R., & Wagner, A. (2022). The impact of mortality shocks on modelling and insurance valuation as exemplified by COVID-19. *Annals of Actuarial Science*, *16*(3), 498–526. https://doi.org/10.1017/S1748499522000045

Ševčíková, H., Li, N., Kantorová, V., Gerland, P., & Raftery, A. E. (2016). Age-specific mortality and fertility rates for probabilistic population projections. In R. Schoen (Ed.), *Dynamic Demographic Analysis* (pp. 285–310). Springer International Publishing. https://doi.org/10.1007/978-3-319-26603-9_15

Ševčíková, H., Raftery, A. E., & Gerland, P. (2018). Probabilistic projection of subnational total fertility rates. *Demographic Research*, *38*, 1843–1884. https://doi.org/10.4054/DemRes.2018.38.60

Ševčíková, H., Raymer, J., & Raftery, A. E. (2024). Forecasting net migration by age: The flow-difference approach. https://doi.org/10.48550/arXiv.2411.09878

Smith, S. K., Tayman, J., & Swanson, D. A. (2013). *A practitioner's guide to state and local population projections*. Springer Netherlands. https://doi.org/10.1007/978-94-007-7551-0

Smith, T. R., & Wakefield, J. (2016). A review and comparison of age–period–cohort models for cancer incidence. *Statistical Science*, *31*(4), 591–610. https://doi.org/10.1214/16-STS580

Stan Development Team. (2024a). RStan: The R interface to Stan [R package version 2.32.6]. https://mc-stan.org/

Stan Development Team. (2024b). Stan functions reference [Version 2.36]. http://mc-stan.org/

Stan Development Team. (2024c). Stan reference manual [Version 2.36]. http://mc-stan.org/

Statistisches Bundesamt. (2022). 15. koordinierte Bevölkerungsvorausberechnung. Retrieved July 3, 2025, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsvorausberechnung/begleitheft.html

Sundberg, L., Agahi, N., Fritzell, J., & Fors, S. (2018). Why is the gender gap in life expectancy decreasing? The impact of age-and cause-specific mortality in Sweden 1997–2014. *International journal of public health*, *63*(6), 673–681. https://doi.org/10.1007/s00038-018-1097-3

Swanson, D. A., & Tayman, J. (2014). Measuring uncertainty in population forecasts: A new approach. In M. Marsili & G. Capacci (Eds.), *Proceedings of the sixth Eurostat/UNECE work session on demographic projections* (pp. 203–215). National Institute of Statistics.

Swanson, D. A., Tayman, J., & Cline, M. (2025). A new approach to probabilistic county population forecasting with an example application to West Texas. *Population*

Research and Policy Review, *44*(4), 43. https://doi.org/10.1007/s11113-025-09961-3

Tayman, J. (2011). Assessing uncertainty in small area forecasts: State of the practice and implementation strategy. *Population Research and Policy Review*, *30*(5), 781–800. https://doi.org/10.1007/s11113-011-9210-9

Tibbits, M. M., Groendyke, C., Haran, M., & Liechty, J. C. (2014). Automated factor slice sampling. *Journal of Computational and Graphical Statistics*, *23*(2), 543–563. https://doi.org/10.1080/10618600.2013.791193

Turek, D., de Valpine, P., & Paciorek, C. (2024). nimbleHMC: Hamiltonian Monte Carlo and other gradient-based MCMC sampling algorithms for 'nimble' [Version 0.2.3].

United Nations. (2024). *World population prospects 2024: Methodology of the United Nations population estimates and projections* (Research Report No. UN DESA/POP/2024/DC/NO. 10). Population Division, Departement of Economic and Social Affairs, United Nations. New York.

van Berkum, F., Melenberg, B., & Vellekoop, M. (2025). Estimating the impact of the COVID-19 pandemic using granular mortality data. *Insurance: Mathematics and Economics*, *121*. https://doi.org/10.1016/j.insmatheco.2025.01.001

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., Gelman, A., Goodrich, B., Piironen, J., Nicenboim, B., & Lindgren, L. (2023). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models [Version 2.6.0]. https://cran.r-project.org/web/packages/loo/index.html

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., & Clark, S. J. (2019). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, *28*(9), 2614–2634. https://doi.org/10.1177/0962280218767988

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. John Wiley & Sons. https://doi.org/10.1002/0471662682.fmatter

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12), 3571–3594.

Welch, N. G., Ševčíková, H., & Raftery, A. E. (2024). Bringing age back in: Accounting for population age distribution in forecasting migration. *arXiv preprint arXiv:2403.05566*.

Wilson, T. (2012). Forecast accuracy and uncertainty of Australian Bureau of Statistics State and Territory population projections. *International Journal of Population Research*, *2012*(1). https://doi.org/10.1155/2012/419824

Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2022). Methods for small area population forecasts: State-of-the-art and research needs. *Population Research and Policy Review*, *41*(3), 865–898. https://doi.org/10.1007/s11113-021-09671-6

Wiśniowski, A., & Raymer, J. (2025). Multiregional population forecasting: A unifying probabilistic approach for modelling the components of change. *European Journal of Population*, *41*(1), 11. https://doi.org/10.1007/s10680-025-09729-7

Wiśniowski, A., Smith, P. W. F., Bijak, J., Raymer, J., & Forster, J. J. (2015). Bayesian population forecasting: Extending the Lee-Carter method. *Demography*, *52*(3), 1035–1059. https://doi.org/10.1007/s13524-015-0389-y

Wolff, M., Haase, A., Leibert, T., & Cunningham Sabot, E. (2022). Calm ocean or stormy sea? Tracing 30 years of demographic spatial development in Germany. *Cybergeo: European Journal of Geography*. https://doi.org/10.4000/cybergeo.38031

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wong, J. S. T., Forster, J. J., & Smith, P. W. F. (2023). Bayesian model comparison for mortality forecasting. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *72*(3), 566–586. https://doi.org/10.1093/jrsssc/qlad021

Wong, J. S., Forster, J. J., & Smith, P. W. (2018). Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics*, *83*, 206–221. https://doi.org/10.1016/j.insmatheco.2017.09.023

Xu, H., Logan, J. R., & Short, S. E. (2014). Integrating space with place in health research: A multilevel spatial investigation using child mortality in 1880 Newark, New Jersey. *Demography*, *51*(3), 811–834. https://doi.org/10.1007/s13524-014-0292-y

Yang, Y., Shang, H. L., & Cohen, J. E. (2022). Temporal and spatial Taylor's Law: Application to Japanese subnational mortality rates. *Journal of the Royal Statistical*

Society Series A: Statistics in Society, *185*(4), 1979–2006. https://doi.org/10.1111/rssa.12859

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007. https://doi.org/10.1214/17-BA1091

Yeung, J., Alexander, M., & Riffe, T. (2023). Bayesian implementation of Rogers–Castro model migration schedules: An alternative technique for parameter estimation. *Demographic Research*, *49*, 1201–1228. https://doi.org/10.4054/DemRes.2023.49.42

Yu, C. C., Ševčíková, H., Raftery, A. E., & Curran, S. R. (2023). Probabilistic county-level population projections. *Demography*, *60*(3), 915–937. https://doi.org/10.1215/00703370-10772782

Zhang, J. L., & Bryant, J. (2020). Bayesian disaggregated forecasts: Internal migration in Iceland. In S. Mazzuco & N. Keilman (Eds.), *Developments in demographic forecasting* (pp. 193–215). Dordrecht, The Netherlands: Springer.

Zhou, R., & Li, J. S.-H. (2022). A multi-parameter-level model for simulating future mortality scenarios with COVID-alike effects. *Annals of Actuarial Science*, *16*(3), 453–477. https://doi.org/10.1017/S1748499522000033