# Bayesian Forecasting of Mortality Rates for Small Areas Using Spatiotemporal Models

Julius Goes*

*University of Bamberg, Institute of Statistics*

Version: Accepted Manuscript

**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

*julius.goes@uni-bamberg.de

## 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 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, the an application of the TOPLAS relational model (Rau & Schmertmann 2020; Schmertmann & Gonzaga 2018), as well as more recently using Taylor's law 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 our 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.

## 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 Statistical Institute. The datasets consist of the total number of deaths (Bayerisches Landesamt für Statisitk 2022b) as well as population counts (Bayerisches Landesamt für Statisitk 2022a) 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 - 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.

## 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 $x = 1, ..., X$ at time $t = 1, ..., T$ in region $r = 1, ..., R$ with $\boldsymbol{y} = (y_{1,1,1}, ..., y_{X,T,R})^\intercal$. Moreover, assume that $y_{x,t,r}|M_{x,t,r} \sim \mathsf{Poi}(E_{x,t,r}M_{x,t,r})$, 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. 2000:15). 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 a follows

$$\eta_{x,t,r} = \mu + \alpha_x + \kappa_t + \gamma_k + \phi_r + \varepsilon_{x,t,r}. \tag{1}$$

Here, the parameter $\phi_r$ denotes the regional effect capturing spatial heterogeneity and $\mu$ the global intercept, that is, the average log-mortality rate. Parameters $\alpha_x$, $\kappa_t$ and $\gamma_k$ denote the respective age, time and cohort effects. The age effect is a static function describing age related effects, while $\kappa_t$ denotes the evolution of mortality over time. The cohort effect $\gamma_k$ 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,

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

Here, the parameter $\alpha_x$ denotes the average log-mortality rate at age $x$. This static age function models the general shape of mortality by age. The parameter $\kappa_t$estimates the global change over time and $\gamma_k$, the global effect of cohort $k$. The parameters $\beta_x^{(1)}$ and $\beta_x^{(2)}$ measure the response to changes of $\kappa_t$ and $\gamma_k$, respectively at age $x$. The regional effect $\phi_r$ 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, we invoke the typical constraints, that is $\sum_x \beta_x^{(1)} = \sum_x \beta_x^{(2)} = 1$ and $\sum_t \kappa_t = \sum_k \gamma_k = 0$ (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 $\mu$ to be identifiable. This can be achieved by either a corner constraint, that

is, to fix a parameter value to (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) demonstrate that this depends on the way future period and cohort effect are extrapolated. Examples include the random walk with drift and random walk of order two. Alternatively, a zero mean forecast may be used (Smith & Wakefield 2016).

### 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 zero.[1] Hence, it does not make sense to set an uninformative prior for the intercept, like $\mu \sim \mathcal{N}(0, 10000)$, 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

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

[1] 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)

with $\nu_x \overset{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$. Which 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 $\sigma_\alpha^2$. 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 $i = 1, 2$.

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

$$\kappa_t = c + \kappa_{t-1} + \omega_t, \tag{3}$$

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 (that is 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 $\boldsymbol{u} = (u_1, \ldots, u_R)^\intercal$ 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

$$p(u_r|u_{-r}; \sigma_u^2) \sim \mathcal{N}\left(\frac{\sum_{r \neq j} w_{rj} u_j}{\sum_{r \neq j} w_{rj}}, \frac{\sigma_u^2}{\sum_{r \neq j} w_{rj}}\right), \tag{4}$$

where $\sigma_u^2$ 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 effect $u_r$ is normally distributed with mean equal to the average of its neighbors. It should be noted, that the distribution in Eq. (4) is improper as it only defines the differences between pairs of units and not its overall level. This distribution cannot be used as a model for the data but still serve as a prior (Banerjee et al. 2015:155). The spatially structured effect is therefore constrained to sum to 0, that is $\sum_r u_r = 0$.

The ICAR prior of Eq. (4) does not allow for spatially unstructured variation. To address this issue, the BYM2 model combines both a spatially structured and unstructured component. In addition, a single variance parameter $\sigma_\phi^2$ is placed on the combined components, while a mixing parameter $\rho \in [0, 1]$ accounts for the amount spatial or non spatial variation. Let $\boldsymbol{\phi} = (\phi_1, ..., \phi_R)^\intercal$, then the BYM2 prior is given by

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

Here, $\boldsymbol{v}^* = (v_1^*, ..., v_R^*)^\intercal$ 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 one (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 $\rho$ is given a Beta$(0.5, 0.5)$ hyperprior.

**Forecasting**

In a Bayesian setting, forecasting is part of the estimation process and carried out via evaluation of the posterior predictive distribution. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \mu, c)^\intercal$ 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 $s$th $(s = 1, \ldots, S)$

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. (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 $s$-th posterior draw of all remaining parameters and plug all values into Eq. (2), respectively Eq. (1) to compute $\log(\hat{M}_{x,T+h,r}^{(s)})$.

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

After having obtained forecasts of mortality rates $\hat{M}_{x,T+h,r}$ they can be converted into deaths $\hat{y}_{x,T+h,r}$. For each posterior predictive draw, sample from a Poisson distribution to obtain $\hat{y}_{x,T+h,r}^{(s)}$, with $\hat{y}_{x,T+h,r}^{(s)} \sim \mathsf{Poi}\left(E_{x,T+h,r}\hat{M}_{x,T+h,r}^{(s)}\right)$.. 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.

## Estimation of Parameters

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 2023b), 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 2023a). 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 is available at: https://github.com/goesj/BavarianMortality. Information regarding the size of the chains and number of iterations may be found in the Appendix.

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 non-convergence 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, we refer to Vehtari et al. (2021). The authors argue for a very tight threshold and propose to only 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 ESS, 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 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..

## 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 $\mathcal{M} = (M_1, ..., M_k)$ 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, though 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

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

where $\mathcal{D}$ denotes a suitable scoring function, for example, the negative log score [2] and

$$\mathcal{A}^K = \Big\{ a_k \in [0,1]^K : \sum_{k=1}^{K} a_k = 1 \Big\}.$$

The stacked $h$- step ahead predictive distribution is then

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

## 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.

### 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. (2015:75-76) 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 (2004:235) propose to replace the death counts

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

$y_{x,t,r}$ by a standardized value $z_{x,t,r}$ under the constant risk hypothesis

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

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

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

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 Fig. F.1.

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.

### 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 $\boldsymbol{y_{1:J}} = (y_{1,1,1}, ..., y_{X,J,R})^\intercal$. The remaining $L = T - J$ years are considered as test set, with $l = J + 1, ..., J + L$.

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 $\overline{\widehat{M}}_{x,l,r}$ 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

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

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

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. It is 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. 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

$$P(y_l \leq \hat{Y}_{T+h} \leq y_u) \geq p_{Cov}. \tag{5}$$

Standard Bayesian practice computes PIs by calculating sample quantiles of the posterior predictive distribution, and then use 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. (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, we can still aim to find PIs with the desired coverage level $p_{Cov}$. Homburg et al. (2021) introduce 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. (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 PI's based on empirical quantiles in a small simulation study we created. Therefore, PIs for forecasted deaths are calculated using the algorithm of Homburg et al. (2021). Details about the algorithm and said study can be found in the Appendix.

## 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.

**Table 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 |

### 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 Fig. F.2, respectively Fig. F.3. 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 Fig. F.4 and Fig. F.5 in the Appendix for females and males. I therefore conclude, that both the RH_BYM2 and APC_BYM2 are suitable at explaining the data.

### 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.

*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 2. 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 2016-2017, the APC_BYM2 clearly outperforms the RH_BYM2 Results are given in Table E.1 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 - 2017. The results are found in Table 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 Table E.1 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 as well as 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, a satisfactory mark.

**Table 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.

*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 3.

**Table 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.

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 3), an unsurprising result considering the performance of the RH_BYM2 model in the test period. Results split by years may be found in Table E.2 in the Appendix. Coverage at the 80% level is a little higher than the nominal value of 80% for all models, similar to the female data.

The results for males are interesting in the sense that the data at hand seem to exhibit

**Table 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.

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

**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 in Table 4.

The final weights for the female data are the following: .906 for the APC_BYM2 and .094 for the RH_BYM2 model. These are similar weights to those received in the model evaluation period with the weight of the RH_BYM2 decreasing a bit. For males, the final weights are as follows: .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 shows the mean LE estimates of all regions split by sex. Panel a displays those of females while 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 Statisitk (2020:13). 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 life expectancy (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 Fig. F.7 for females and Fig. F.8 for males. The plots are sorted by mean LE.

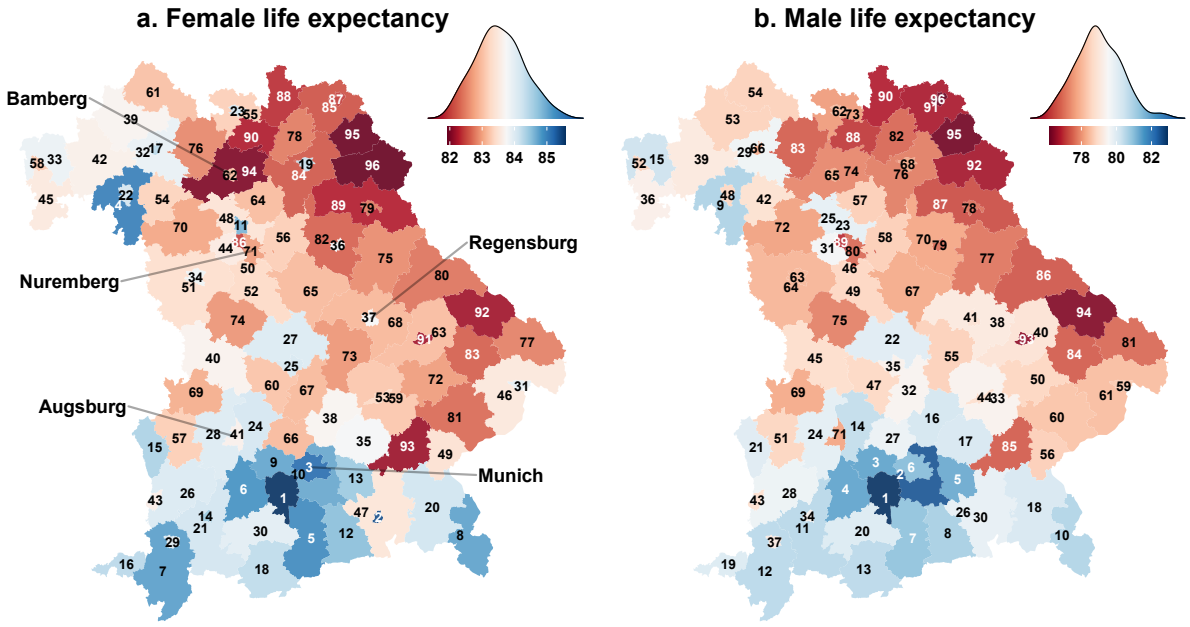


**Figure 1:** Mean estimates of life expectancy for all regions of Bavaria in 2017, using the stacking approach. Female estimates are plotted in panel a, and values for males are shown in panel b. Red areas denote regions with lower life expectancy, while blue areas denote regions with higher life expectancy. 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.

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 Fig. 2. The complete results for all other regions is available in a Shiny app online. [3]

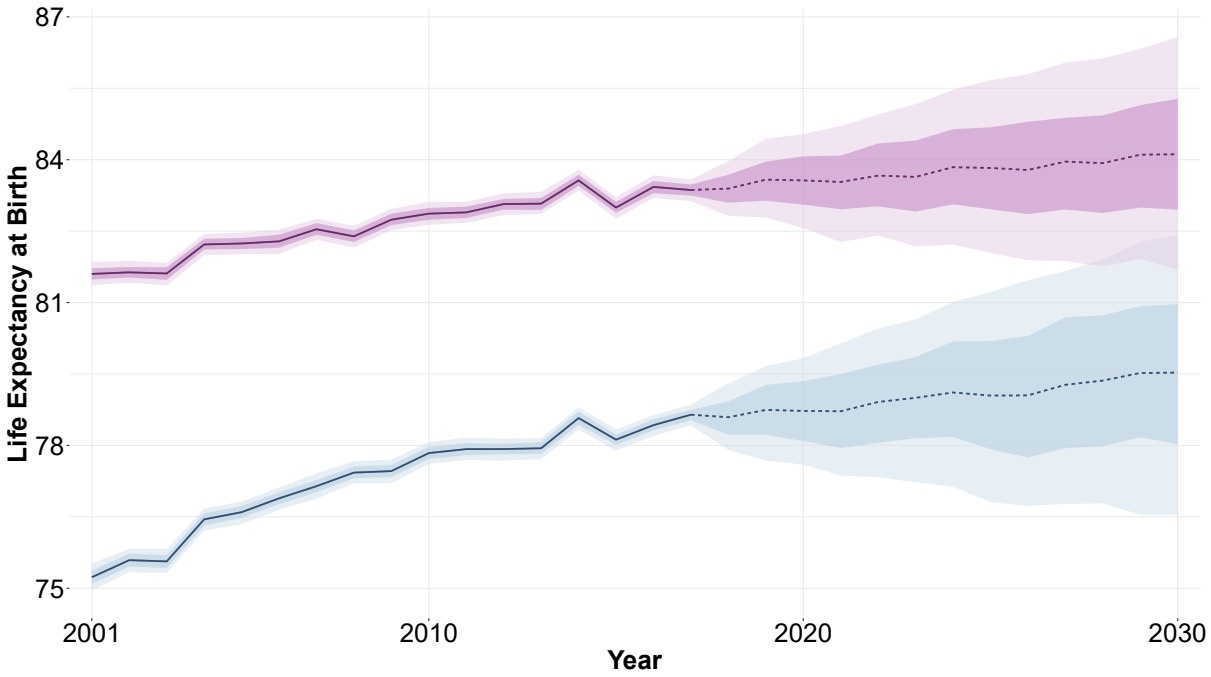


**Figure 2:** Prediction and estimation of life expectancy for the region of Bamberg, City, using the stacking approach. The solid lines denote the mean estimation for 2001 to 2017, and the dashed lines denote the mean prediction for 2018 to 2030. The shades of purple and blue represent PI for females and males respectively. Darker shades show 50% PIs, while lighter shades show 80% PIs.

Additionally, when examining Fig. 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 found in Fig. 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 men.

Lastly, Fig. 2 shows a relatively sharp increase in LE is seen 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.

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

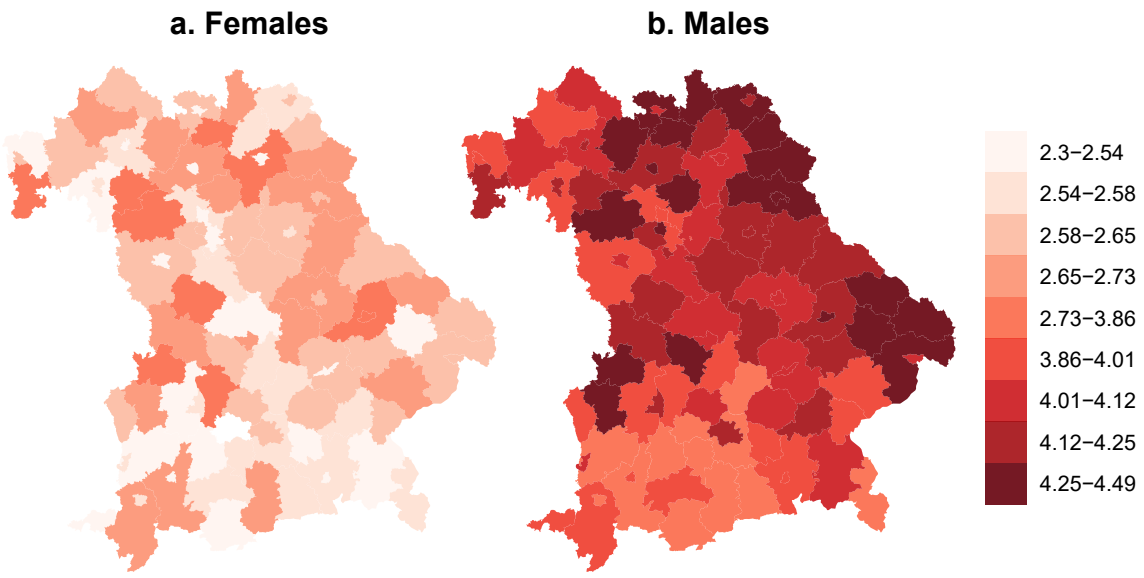


**Figure 3:** Differences in mean life expectancy between 2001 and 2030 for all regions in Bavaria, by sex, using the stacking approach. Panel a shows the difference for females and panel b shows that for males. Darker shades indicate regions with a greater increase in life expectancy.

## 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) incorporates 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.

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

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 our 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) introduces single period jumps into the LC model. These jumps may be incorporated into our model framework as well and is potential for further research.

## 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 looking at the forecasted death rates we notice slightly lower values in terms of scoring rules and MAE for females compared with those of males (see Table 2, respectively Table 3). 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

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

## References


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*, 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.

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

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

Bayerisches Landesamt für Statisitk. (2020). Bruttoinlandsprodukt und bruttowertschöpfung in bayern 2012 bis 2020 [gross domestic product and gross value added in bavaria 2012 to 2020]. *Statistical Reports*, (202000). https://www.statistik.bayern.de/mam/produkte/veroffentlichungen/statistische_berichte/p1300c_202000.pdf

Bayerisches Landesamt für Statisitk. (2022a). 12411-007s. *GENESIS-Online: CC BY 3.0 DE [Data set]*. https://www.statistikdaten.bayern.de/genesis/online/logon.

Bayerisches Landesamt für Statisitk. (2022b). 12613-108s. *GENESIS-Online: CC BY 3.0 DE [Data set]*. https://www.statistikdaten.bayern.de/genesis/online/logon.

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.

Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, *70*(1), 1–19.

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

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

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*(4), 1337–1363.

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.

Cairns, A. J., 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.

Congdon, P. (2014). Estimating life expectancies for US small areas: A regression framework. *Journal of Geographical Systems*, *16*(1), 1–18.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, *65*(4), 1254–1261.

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

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

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*(3), 115–146.

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

Heuer, C. (1997). Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, *53*, 161–177.

Hobcraft, J., Menken, J., & Preston, S. (1982). Age, period, and cohort effects in demography: A review. *Population Index*, 4–43.

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.

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*, 1–37. https://doi.org/10.18637/jss.v090.i12

Keilman, N. (2020). Evaluating probabilistic population forecasts. *Economie et Statistique / Economics and Statistics*, *520-521*, 49–94. https://doi.org/10.24187/ecostat.2020.520d.2033

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., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, *87*(419), 659–671.

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.

Lorentzen, P., McMillan, J., & Wacziarg, R. (2008). Death and development. *Journal of Economic Growth*, *13*(2), 81–124.

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.

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. *Annals of Applied Statistics*, *9*(4), 1889.

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

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.

Ocaña-Riola, R., & Mayoral-Cortés, J. M. (2010). Spatio-temporal trends of mortality in small areas of southern Spain. *BMC Public Health*, *10*(1), 1–12.

Pedroza, C. (2006). A Bayesian forecasting model: Predicting US male mortality. *Biostatistics*, *7*(4), 530–550.

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

Rau, R., & Schmertmann, C. P. (2020). District-level life expectancy in Germany. *Deutsches Ärzteblatt International*, *117*(29-30), 493. https://doi.org/10.3238/arztebl.2020.0493

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

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.

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.

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

Smith, T. R., & Wakefield, J. (2016). A review and comparison of age–period–cohort models for cancer incidence. *Statistical Science*, *31*(4), 591–610.

Stan Development Team. (2023a). RStan: The R interface to Stan [R package version 2.26.16]. https://mc-stan.org/

Stan Development Team. (2023b). Stan modeling language users guide and reference manual [2.32]. http://mc-stan.org/

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

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.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2), 667–718. 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.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. Hoboken, New Yersey: John Wiley & Sons.

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*, 865–898.

Wiśniowski, A., Smith, P. W., 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

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259.

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., & Cohen, J. (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*, 1979–2006.

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

Zhang, J. L., & Bryant, J. (2020). Bayesian disaggregated forecasts: Internal migration in Iceland. In S. Mazzuco & N. Keilman (Eds.), *The springer series on demographic methods and population analysis: Vol. 49. developments in demographic forecasting* (pp. 193–215). Dordrecht, The Netherlands: Springer.

# Appendix

## A 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.

## B Derivation of DSS Parameters

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

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

**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 |

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

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

# C 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 $P_m \sim \mathsf{Poi}(\lambda_m)$, 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 $\hat{P}_m \sim \mathsf{Poi}(\lambda_m)$
    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 $\{c_1, \ldots, c_M\}$
**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. C.1.

# D 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. F.1. Values diverting from one show spatial association of some form. Fig. F.1 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.

## D.1 Posterior Predictive Checks

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

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

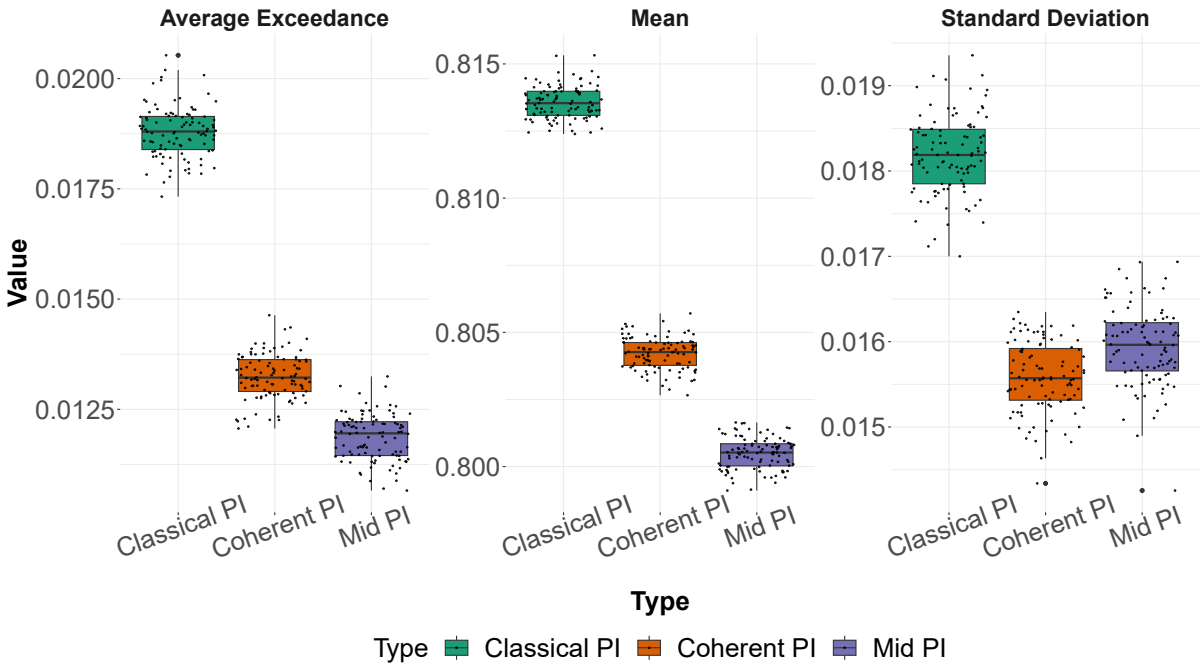


**Figure C.1:** Boxplot with results of the simulation study. The small black points denote the results of each sample statistic for each iteration by PIs type, while the boxplot is a summary of those points.

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. F.2, respectively Fig. F.3.

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. F.4, respectively Fig. F.5.

## D.2 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) 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. F.6).

# E Additional Tables

**Table E.1:** 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 E.2:** 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 |

# F Additional Figures

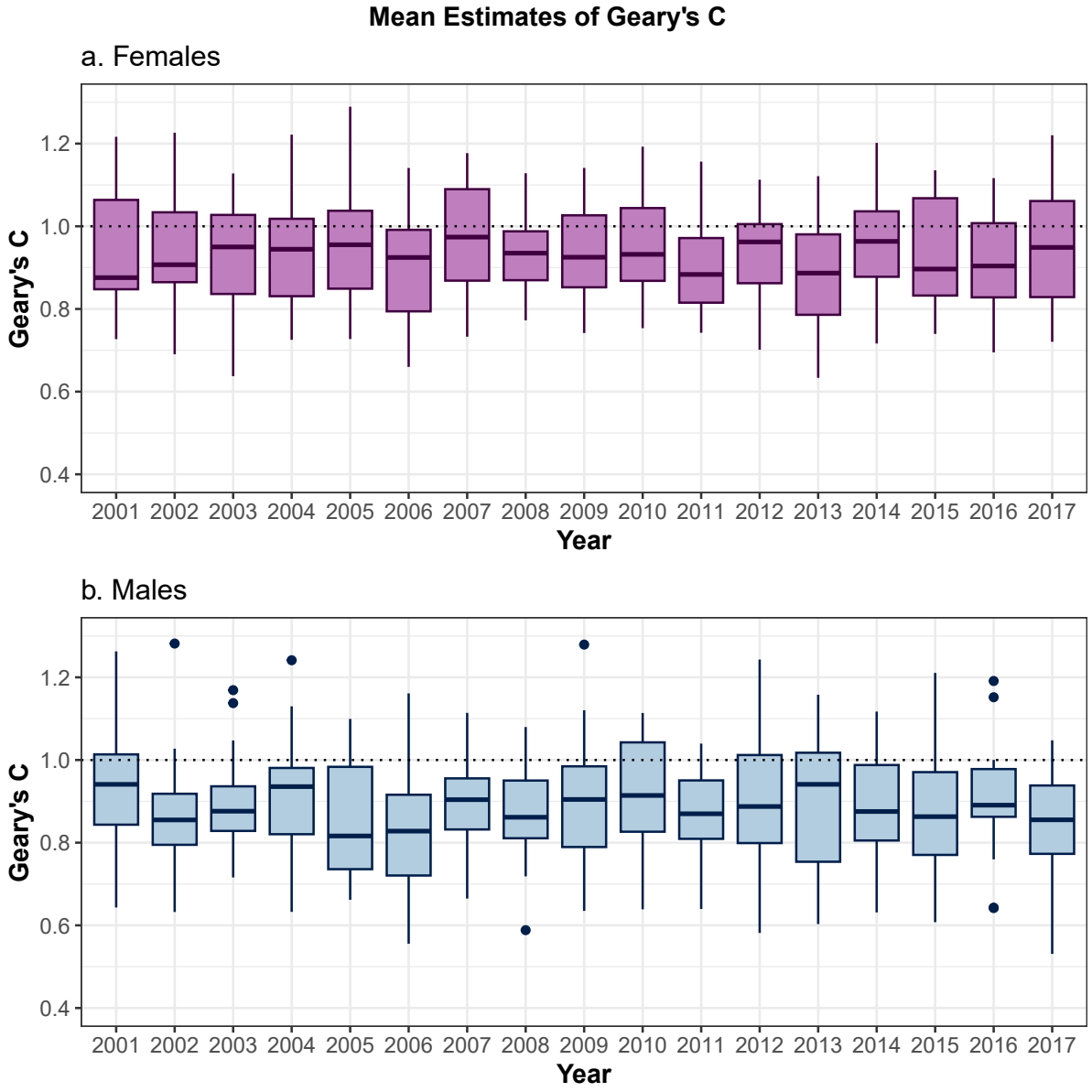


**Figure F.1:** The distribution of Geary's $C$ values over all age groups by year. Estimates are given for both females (panel a) and males (panel b). The dashed line at 1 denotes the value of Geary's $C$ under spatial independence.

## F.1 Posterior Predictive Checks

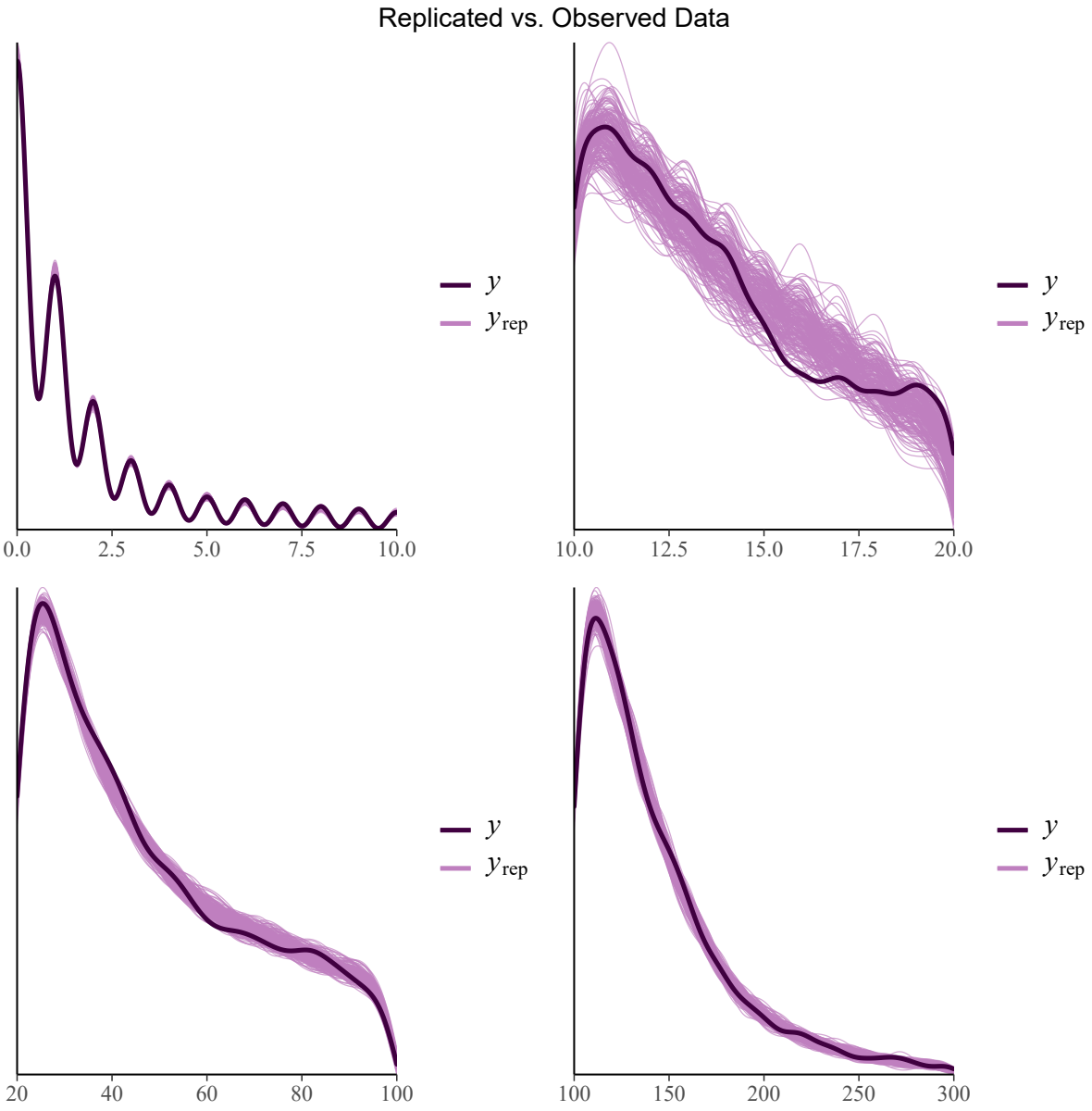


**Figure F.2:** Replicated vs. observed density of in-sample-values for females . The x-axis shows the actual value while the y-axis the respective density. Thick darker line denotes observed, while each light line denotes a realization of the replicated density. The plot is split along the x-axis for a better overview.

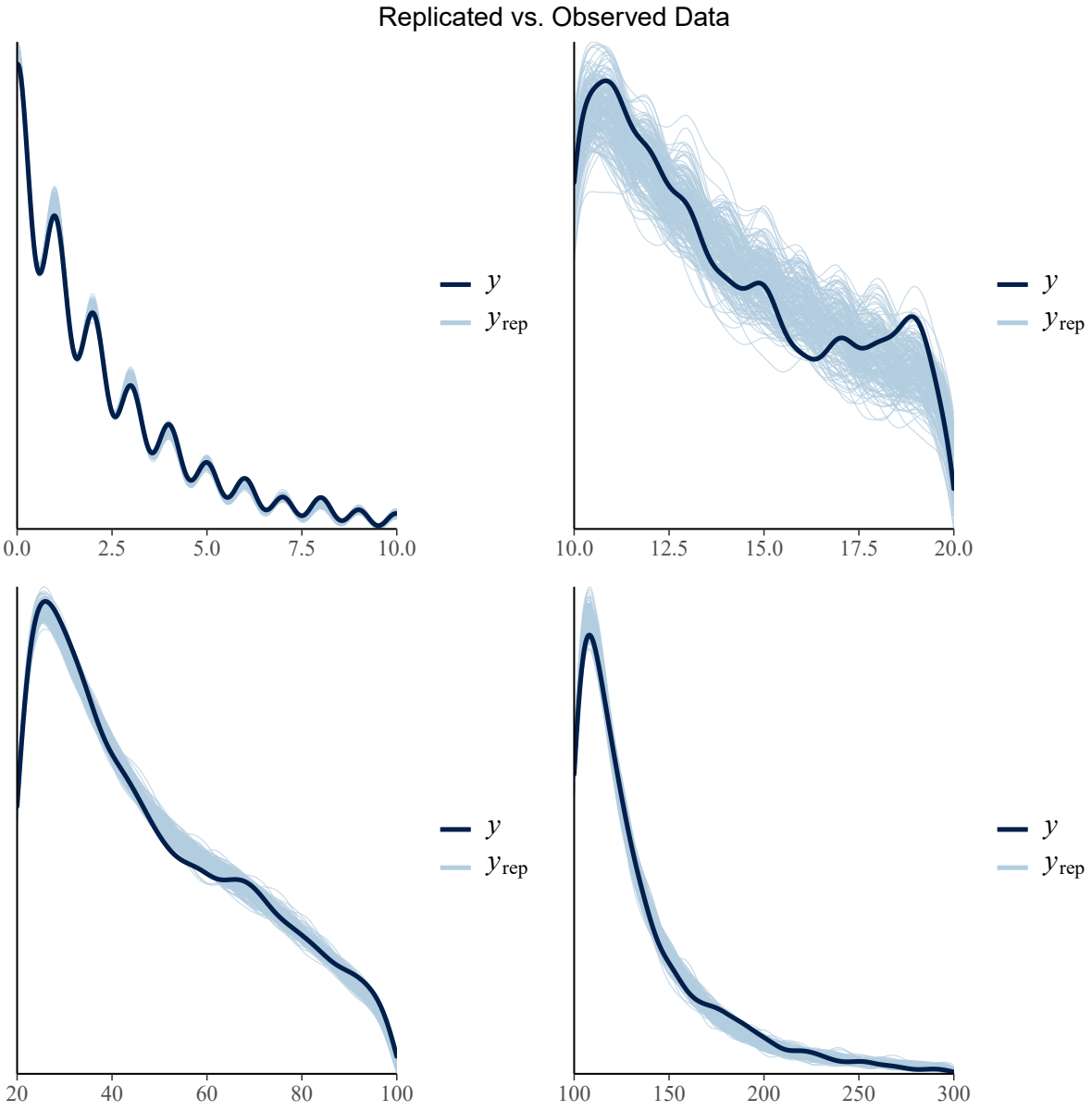


**Figure F.3:** Replicated vs. observed density of in-sample values for males. The x-axis shows the actual value while the y-axis the respective density. Thick darker line denotes observed, while each light line denotes a realization of the replicated density. The plot is split along the x-axis for a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

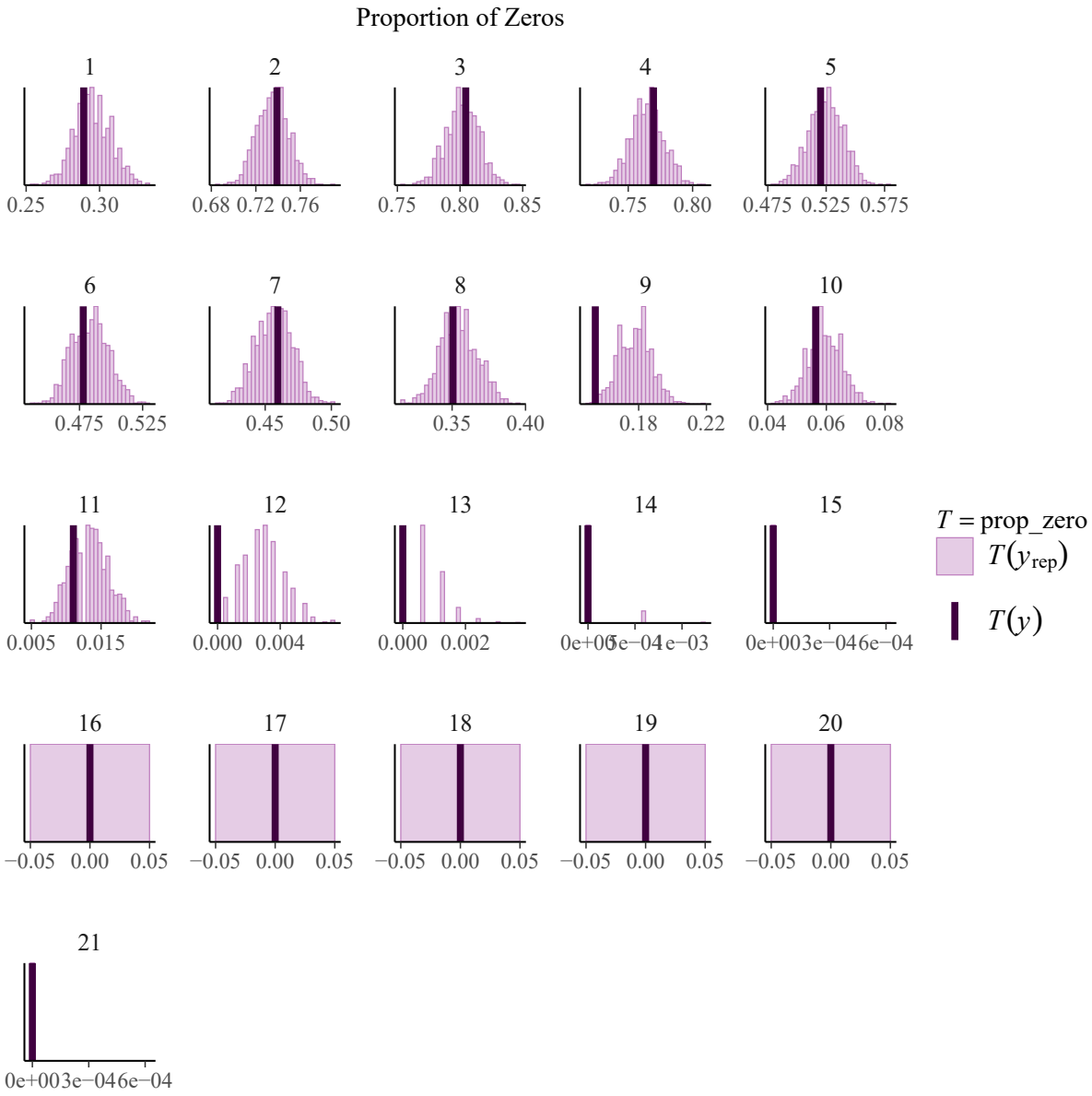


**Figure F.4:** Histogram of the test statistic $T[\boldsymbol{y}_{\boldsymbol{rep}}^{(\boldsymbol{x})}, \boldsymbol{\theta}]$ proportion of zeros, for the replicated data set. The observed data test statistic is given by the vertical line for all age groups of the APC_BYM2 for females.

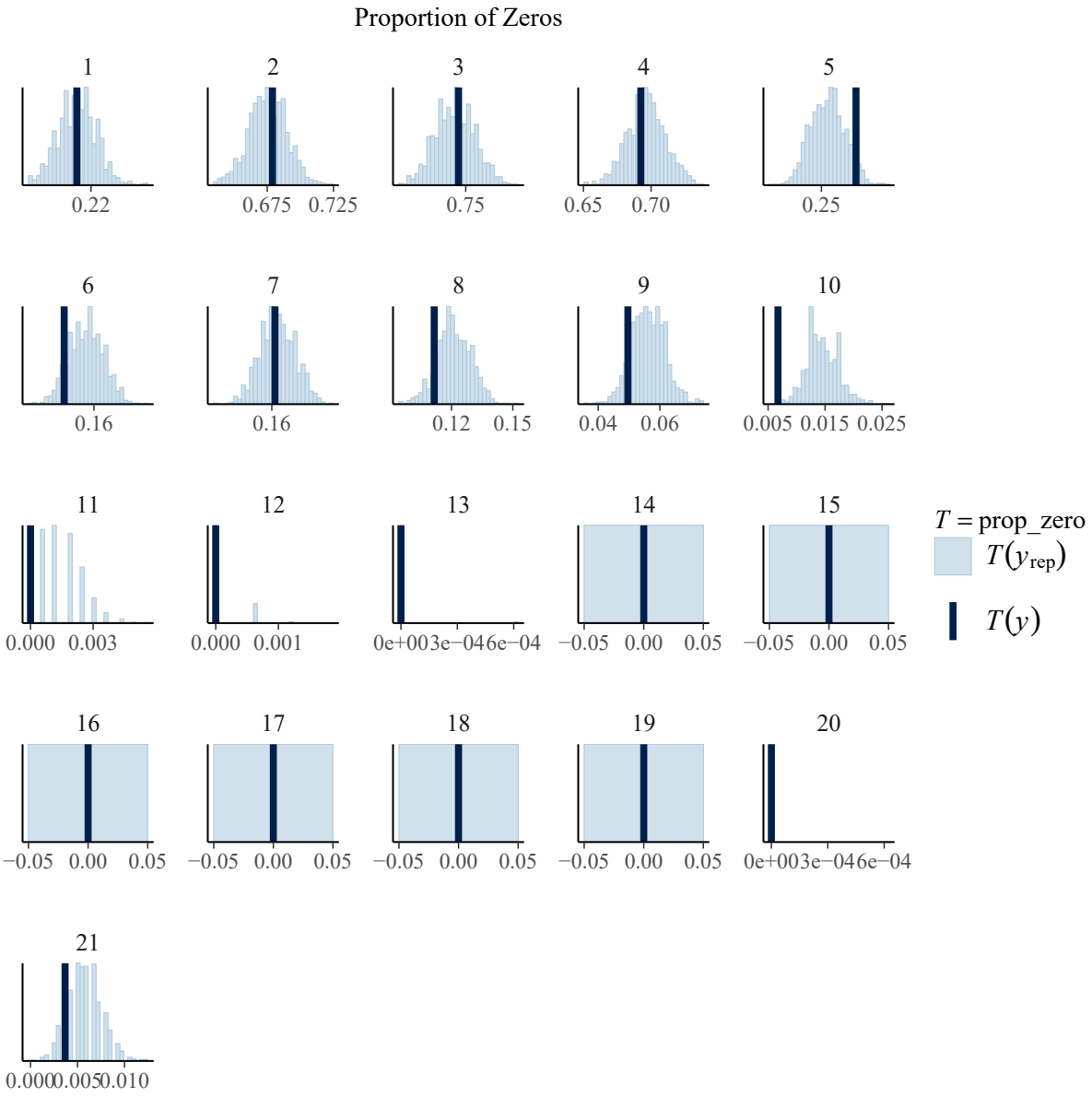


**Figure F.5:** Histogram of the test statistic $T[\boldsymbol{y}_{\boldsymbol{rep}}^{(\boldsymbol{x})}, \boldsymbol{\theta}]$ proportion of zeros, for the replicated data set. The observed data test statistic is given by the vertical line for all age groups of the APC_BYM2 for males.

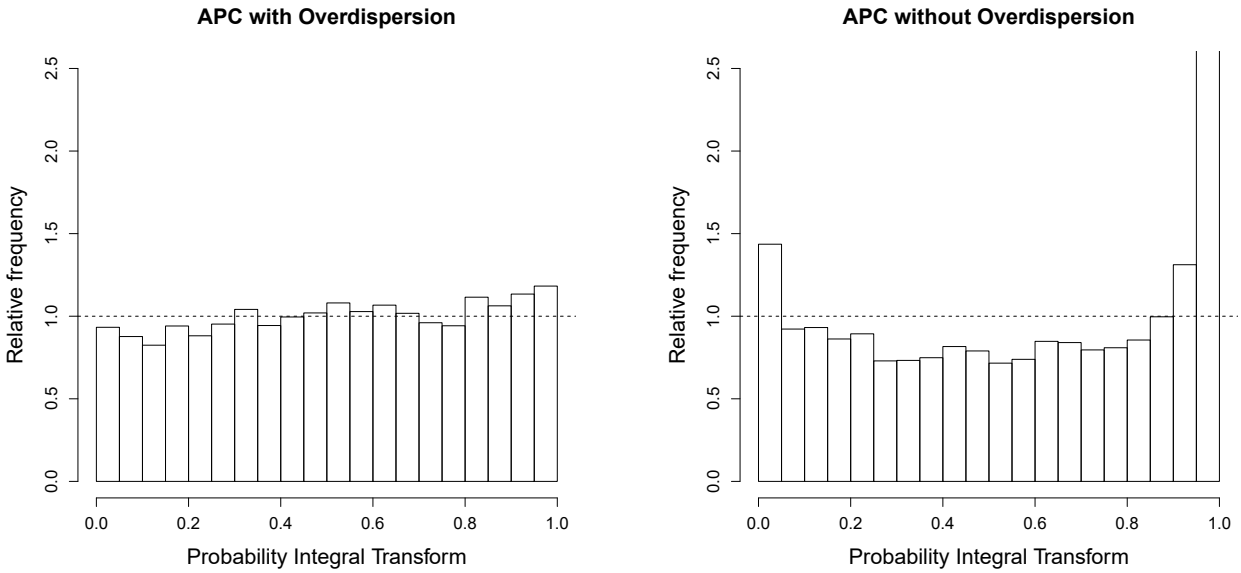


**Figure F.6:** Nonrandomized PIT for probabilistic forecast of APC Model for females with overdispersion parameter $\varepsilon_{x,t,r}$ (left) compared with probabilistic forecast of frequentistic APC Model without overdispersion parameter (right).

## F.2 Estimated Life Expectancy

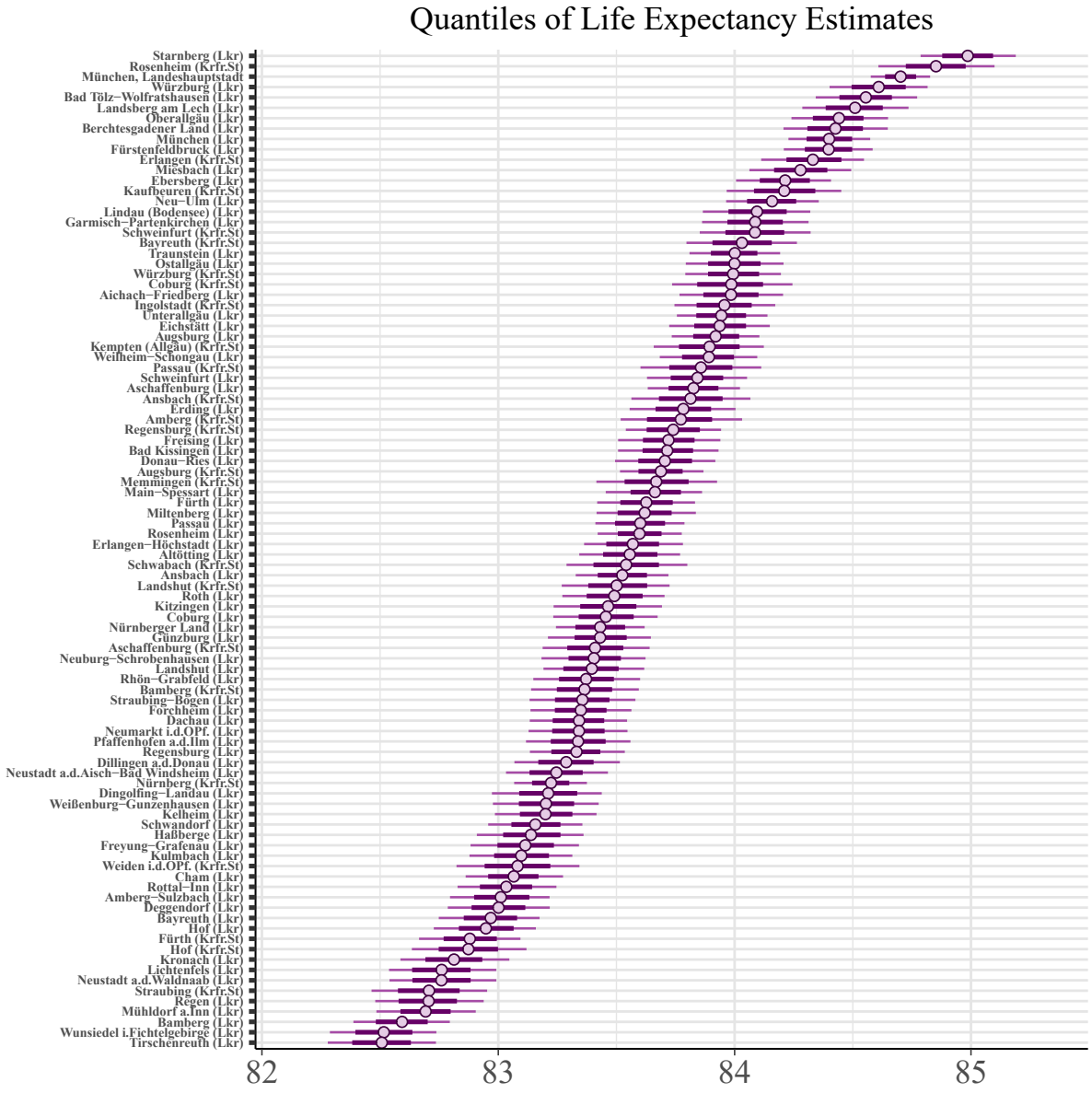


**Figure F.7:** Female life expectancy estimates in 2017. Points denote the mean value. The thick bar denotes the 50% while the thin the 80%- prediction interval

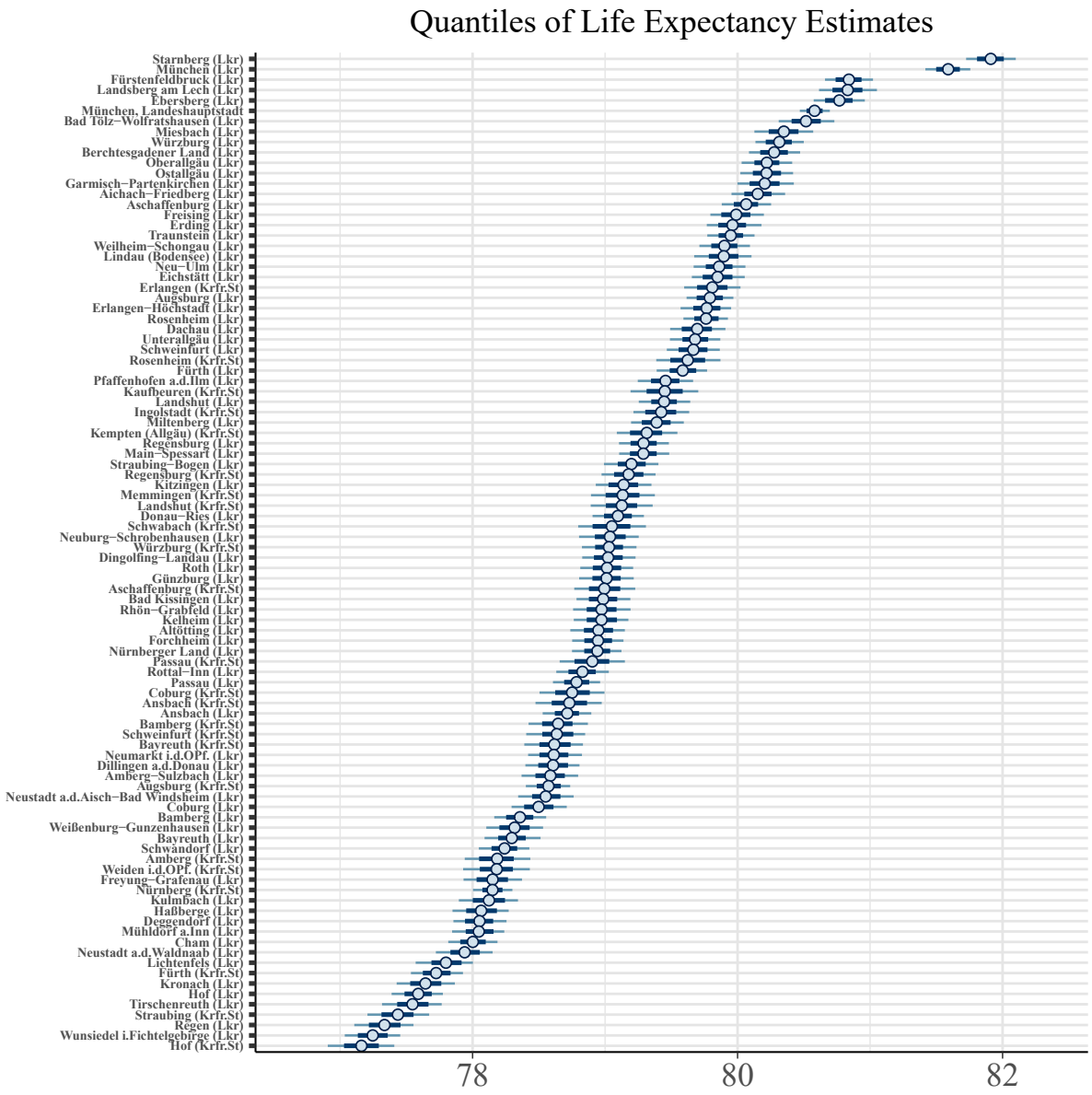


**Figure F.8:** Male life expectancy estimates in 2017. Points denote the mean value. The thick bar denotes the 50% while the thin the 80%- prediction interval