Mortality and Longevity

Aging and Retirement

Multi-Population Longevity Models: A Spatial Random Field Approach

January 2020

Multi-Population Longevity Models: A Spatial Random Field Approach

AUTHOR

Nhan Huynh
University of California at Santa Barbara

Mike Ludkovski, Ph.D.
University of California at Santa Barbara

Howard Zail
Elucider, LLC

SPONSOR

Society of Actuaries:
Aging and Retirement Research
Committee on Life Insurance Research
Financial Reporting Section
Mortality and Longevity Research
Product Development Section

Caveat and Disclaimer

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information.

Multi-Population Longevity Models: a Spatial Random Field Approach

Nhan Huynh[*], Mike Ludkovski[†]and Howard Zail[‡]

September 27, 2019

Abstract

We investigate joint modeling of longevity trends using the spatial statistical framework of Gaussian Process regression. Our analysis is motivated by considering the Human Mortality Database that provides raw mortality tables for nearly 40 countries and clearly demonstrates the commonality in global longevity. Yet few stochastic models exist for handling more than two populations at a time. To bridge this gap, we develop a spatial covariance approach that treats mortality data through the lens of smoothing and forecasting noisy input-output relationships. In our framework, multiple populations are approached as distinct levels of a factor covariate, explicitly capturing the cross-population dependence. We demonstrate that our approach not only provides improved accuracy, but intrinsically generates coherent joint future longevity scenarios. It also offers an opportunity to borrow the most recently available data from other datasets, leading to more precise (and statistically more credible) forecasts regarding mortality improvement rates. All the numerical algorithms are implemented using `R` and `Stan` statistical languages and are publicly available. We illustrate using numerous figures on multiple European HMD datasets for both Males and Females.

1 Mortality Models across Multiple Populations

Mortality data are typically collected by jurisdictional areas, such as countries and states. As a result global mortality experience is summarized in dozens of national and sub-national registries, presenting a major data-analysis challenge. The burgeoning Human Mortality Database (HMD 2018) offers a centralized portal to nearly 40 such datasets, yielding a rich source of cross-national longevity trends.

Significant value can be extracted from joint models of these mortality tables. By aggregating data, one hopes to improve prediction accuracy (through better disentangling of trends and "noise") and simultaneously reduce model risk (by increasing credibility of the forecasts). Moreover, joint models allow information fusion, which is very valuable since mortality data are released asynchronously. With a joint model one can rely on the newly released data of a related foreign population to update and improve the domestic forecast. Last but not least, joint models are critical for generating forecasts and future scenarios simultaneously across multiple populations. Individual models will tend to be non-coherent, i.e. include scenarios where the joint mortality trends cross-over or diverge in unrealistic ways.

[*] Department of Statistics and Applied Probability, University of California at Santa Barbara

[†] Department of Statistics and Applied Probability, University of California at Santa Barbara, `ludkovski@pstat.ucsb.edu`

[‡] Elucidor, LLC

Yet few models exist for multi-population longevity analysis besides the 2-population case. The latter case affords the convenient hierarchy of treating one population as the baseline “index” and then separately modeling the “spread” or basis between the index and the secondary population. With three or more populations one may still view one of them as the index, but the conceptual meaning of the multiple resulting longevity spreads becomes fuzzy. Moreover, in the commonly adopted Age-Period-Cohort-style models, multiple populations are treated through decomposition into global- and population-specific factors, implying that the number of factors grows linearly in the number of populations. Since each factor (Age, Period, etc) contains 30+ parameters, one quickly ends up with hundreds of parameters to be estimated, creating significant computation and statistical inference bottlenecks.

To start bridging the gap between the wealth of data in the HMD and multi-population stochastic longevity models, we investigate a *spatial covariance* framework. Our work builds upon the Gaussian Process (GP) models for longevity introduced in Ludkovski et al. (2018). GPs are a machine learning technique that is a centerpiece of probabilistic data science. The main idea of this framework is to view a mortality table as a latent input-output response surface, corrupted by observation noise. Using the Bayesian lens, mortality modeling translates to smoothing (aka interpolating) and extrapolating this surface, specifically using multivariate Gaussian conditioning with respect to observations. This yields a full uncertainty quantification not just for mortality rates, but also for mortality improvement factors.

With a GP approach, extension to multiple populations is conceptually straightforward: we treat populations as a *factor* covariate. The corresponding correlation structure across factor levels (i.e. correlation in mortality experiences of different countries) is then inferred and handled exactly the same as statistical dependence across Ages or Calendar Years. Moreover, we show that GPs are well-suited for all of the joint modeling tasks mentioned above. Their probabilistic structure naturally captures reduced model risk (namely tighter hyperparameter and latent-surface posteriors) and straightforwardly offers borrowing of information from “notched” datasets. Moreover, GPs intrinsically generate coherent and fully-stochastic forecasts.

Related spatio-temporal frameworks were considered by Christiansen et al. (2015) to capture the spread between individual log mortality rates and weighted average log-mortality and Debón et al. (2010). Another related analysis of the HMD can be found in Carracedo et al. (2018) who applied spatio-temporal Markov clustering to detect common patterns of longevity across 26 European countries; see also Antonio et al. (2017). More broadly, there is a growing strand of literature addressing multi-population extensions of the now-classical Lee-Carter stochastic mortality framework. The seminal work by Li and Lee (2005) extended Lee-Carter to two populations, postulating a decomposition of mortality into population-specific plus global Age and Period factors (for a total of $2L + 2$ factors with L populations). More parsimonious versions were proposed by Kleinow (2015) who considered a common Age effect, and Delwarde et al. (2006) who proposed a common Period effect. Enchev et al. (2017) investigated several intermediate cases. Dispensing with country-specific factors allows more interpretability, e.g. in the Kleinow (2015) CAE model one may directly compare period effects across countries since these are scaled with the same age parameters. From the other direction, the model of Li and Lee (2005) functionally corresponds to having a single degree of freedom in the evolution of the mortality curve over time. According to Li (2013) at least two Age/Period factors are warranted, and accordingly a multi-factor extension was investigated. Note that in our setup mortality curves are non-parametric (i.e. as many degrees of freedom as there are data points).

Another way to introduce dependence between populations is through statistical shrinkage within a Bayesian hierarchical model. Raftery et al. (2012) modeled mortality of 160+ countries by first imposing a global hyper-prior over several one-dimensional parameters and then constructing individual Lee-Carter models. A related approach is taken up by Wiśniowski et al. (2015). This framework also permits to inject additional socio-economic or geo-political covariates to capture the varying degree of dependence (Kleinow and Cairns 2013, Boonen and Li 2017).

Several different methods have been proposed to achieve coherent forecasts. In the special situation of two populations, co-integration models are a viable strategy and were studied in Hyndman et al. (2013) for Male-Female mortality; see also the multi-level functional regression approach in Shang (2016). In a similar spirit, D'Amato et al. (2016), Li and Lu (2017), Guibert et al. (2017) investigated Vector Autoregressive (VAR) approaches to achieve correlation across the multiple Period factors of the aforementioned Li and Lee (2005) framework. Alternatively, Chen et al. (2015), Wang et al. (2015) applied a copula approach to capture the dependence between individual Period factors and Yang and Wang (2013) considered a Vector Error Correction model.

The rest of the paper is organized as follows. The next Section 1.1 illustrates the co-dependence of mortality in the context of multiple European nations. Sections 2 and 3 present the GP approach to individual- and multiple-population mortality modeling, respectively. Our primary results are in Section 4; additional features of joint GP models are in Section 5. Finally, Section 6 concludes with the key take-aways and outlook for further analysis.

1.1 Motivation

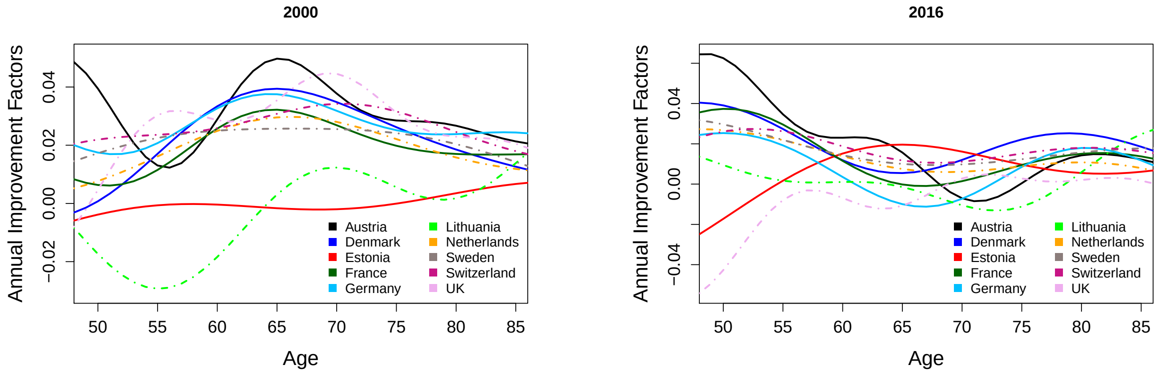
The conceptual driver for joint longevity models is the idea of commonality in mortality experiences of different countries. In other words, there are "global" longevity trends that can be observed across datasets. This similarity is visualized in Figure 1 where we show smoothed mortality improvement rates (i.e. the "gradient" of mortality rate) across 10 European countries: Austria, Denmark, Estonia, France, Germany, Lithuania, Netherlands, Sweden, Switzerland, and UK, see Figure 2.a. We see that there is a strong common pattern, for example in 2016 there is a "wave" structure in Age where mortality improvement in the 60–70 age range is generally lower than at either younger or older ages. Such similarities imply opportunities to fuse information from existing data; they are moreover structura (driven by shared demographics in Western Europe) and are expected to persist into the future. Thus, it is accepted that the mortality forecasts should converge or remain stationary in the long run perpetuating historical regularities (Booth and Tickle 2008). By construction single-population models feature independent, hence divergent, stochastic mortality factors.

In the context of a spatio-temporal model, the commonality of mortality experience implies that there is an underlying *global* covariance structure. This similarity refers not to the specific mortality (improvement) rates across countries, but to the degree that such rates are correlated among themselves. Capturing this correlation is central to reduce *model risk*, i.e. the mis-specification between the true mortality evolution and the fitted model dynamics that arises due to using limited data to calibrate it. Table 1 lists the hyperparameters of the GP models fitted to each of the 10 individual datasets. We observe that most countries have very similar dependence structures, for example almost the same Age coefficients β_1^{ag}, and $\theta_{ag} \in [6, 14]$. However, there are also a few outliers , such as the Switzerland dataset whose θ_{yr} parameter is relatively very large and implies data under-fitting. The latter is in fact statistical anomalies, i.e. the methodology has difficulty in correctly estimating these hyperparameters. As we will show in Section 5.1, by working with more data, a joint model is able to crystallize the underlying dependence pattern and offers a

Table 1: Fitted hyperparameters of single-population GP models. Training set is Ages 50–84 and Years 1990–2016 for Males. Mean function is $m(x) = \beta_0 + \beta_1^{ag} x_{ag}$. Hyperparameter outliers are indicated with italics.

	Denmark	Estonia	Lithuania	Sweden	UK
β_0	-9.9809	-8.4675	-7.6614	-11.1924	-10.3695
β_1^{ag}	0.0935	0.0788	0.0674	0.1071	0.0976
θ_{ag}	11.3093	*16.7860*	9.7826	13.0261	6.5553
θ_{yr}	7.8445	*3.3310*	*2.3587*	10.3797	5.5229
η^2	0.0445	0.0486	0.0155	0.0365	0.0239
σ^2	2.707×10^{-3}	6.436×10^{-3}	2.936×10^{-3}	1.963×10^{-3}	5.421×10^{-4}
	Austria	France	Germany	Netherlands	Switzerland
β_0	-10.4653	-9.7192	-10.2552	-10.9651	-11.0439
β_1^{ag}	0.0991	0.0890	0.0975	0.1053	0.1050
θ_{ag}	6.5078	9.9525	9.2840	13.7416	10.7675
θ_{yr}	5.7786	8.4953	9.0939	6.5164	*17.0494*
η^2	0.0322	0.0435	0.0370	0.0273	0.0724
σ^2	2.052×10^{-3}	4.069×10^{-4}	8.163×10^{-4}	1.235×10^{-3}	2.602×10^{-3}

methodological tool to guard against under- or over-fitting the data.



(a) **(b)**

Figure 1: Smoothed annual Male mortality improvement factors $\partial m_{back}^{GP}(.; yr)$ (11) in different countries, $yr \in \{2000, 2016\}$. The improvement factors are based on fitting a GP model to each individual dataset, following the methodology in Ludkovski et al. (2018). Inference done separately for each country.

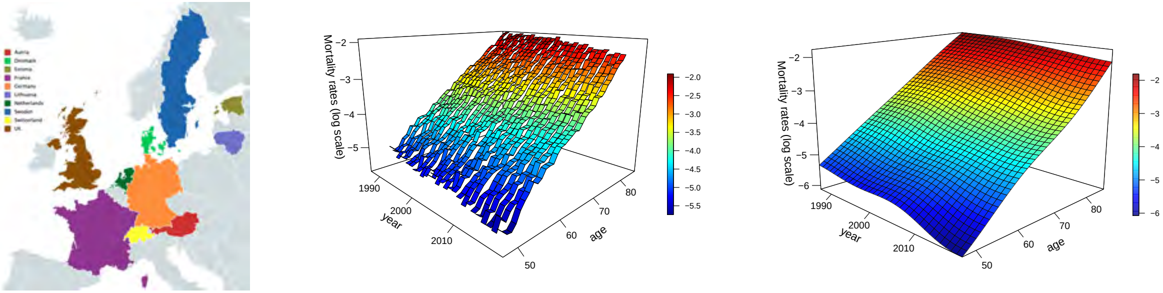
Data Source: We work with mortality data from the Human Mortality Database (HMD) (HMD 2018) which provides the aggregated mortality statistics at the national levels for 40 developed countries across the globe. The HMD applies the same consistent set of procedures (Boe et al. 2015) on each population, which is the reason why mortality data are only available for developed countries whose death registrations and census data are available and reliable. For our analysis we rely on one-year age groups, concentrating on Ages 50–84 for both genders and calendar Years

1990–2016.

The dataset is organized as a large table. The nth observation for the lth country $(l = 1, \ldots, 10)$ contains (i) Age and Year as a pair of independent variables, (x^n_{ag}, x^n_{yr}), and (ii) the logarithm of the observed mortality rate,

$$y^n = \log\left[\frac{\text{Death counts at } (x^n_{ag}, x^n_{yr})}{\text{Exposed-to-risk counts at } (x^n_{ag}, x^n_{yr})}\right] = \log\left[\frac{D^n}{E^n}\right]. \tag{1}$$

We denote by $\mathcal{D}_l = \{(x^n, y^n)\}_{n=1}^N$ the dataset for the lth country. Figure 2.b illustrates a typical mortality surface. It shows the raw Male log mortality rates in Denmark, as well as the smoothed surface obtained from a GP model. We note the prevalent patterns, namely log mortality increasing roughly linearly in Age, and decreasing gradually over calendar Year.



(a) Selected countries

(b) Raw Male log mortality rates in Denmark

(c) Smoothed Male log mortality rates in Denmark

Figure 2: Selected HMD dataset of 10 European countries and an illustration of GP smoothing of mortality observations.

2 Methods

In this section we review the approach of applying Gaussian Process models to individual mortality datasets.

2.1 Gaussian Process Regression for Mortality Tables

A Gaussian process (GP) is an infinite collection of random variables, any finite number of which follows a multivariate normal (MVN) distribution. As such, a GP $f \sim GP(m, C)$ is characterized by its mean function $m(x)$ and its covariance structure $C(x, x')$. This means that for any vector $\mathbf{x} = (x^1, \ldots, x^n)$ of n inputs:

$$f(x^1), \ldots, f(x^n) \sim \mathcal{N}\big(\mathbf{m}(\mathbf{x}), \mathbf{C}(\mathbf{x}, \mathbf{x})\big)$$

where $\mathbf{m}(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ is the mean vector of size n and $\mathbf{C}(\mathbf{x}, \mathbf{x})$ is the n by n covariance matrix, $C(x, x') := \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$.

In a GP regression setup, the latent f links the observations or output vector $\mathbf{y} = (y^1, \ldots, y^n)$ to the input vector $\mathbf{x}$ via:

$$y^i = f(x^i) + \epsilon^i, \tag{2}$$

where ϵ^i is the error term to accommodate the fact that we observe only a noisy sample of $f(x^i)$. In our context, x^i are the individual cells in a mortality table (so indexed by Age, Year, etc.), y^i

are observed raw log mortality rates, and $f(x^i)$ is the *true* mortality rate that would materialize in the absence of any random shocks. We shall assume that the error terms ϵ^i are from i.i.d. Gaussian distribution with zero mean and constant variance across all x's: $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$ or $\epsilon = (\epsilon^1, ..., \epsilon^n) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma} = \text{diag}(\sigma^2))$. It follows that $\mathbf{y} \sim \mathcal{N}(\mathbf{m}(\mathbf{x}), \mathbf{C}(\mathbf{x}, \mathbf{x}) + \mathbf{\Sigma})$, because

$$Cov(y^i, y^j) = Cov(f(x^i), f(x^j)) + \sigma^2 \delta(x^i, x^j) \tag{3}$$

where $\delta(x^i, x^j)$ is the Kronecker delta which is one iff the indices match $i = j$, and zero otherwise.

GP regression works by applying the Bayesian formalism of assigning a prior distribution to $f \sim GP(m, C)$ and using *MVN conditioning* relative to a dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ to infer the posterior distribution of f. The Gaussian structure of the prior and the Gaussian structure of (2) together with Bayes' rule yield a Gaussian posterior for $f|\mathcal{D} \sim GP(m_*, C_*)$:

$$\begin{aligned} \text{Posterior distribution} \ &= \frac{\text{Prior distribution x Likelihood function}}{\text{Marginal distribution}} \\ \text{or } p(\mathbf{f}|\mathcal{D}) &\propto p(\mathbf{f}) p(\mathbf{y}|\mathbf{x}, \Theta). \end{aligned}$$

The principal objective is to draw prediction about $\mathbf{f}_* \equiv f(\mathbf{x}_*)$ or future observations $\mathbf{y}_* \equiv Y(\mathbf{x}_*)$ at new inputs $\mathbf{x}_*$. By construction, $\mathbf{y}$ and $\mathbf{y}_*$ follow a joint MVN distribution:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m}_{**} \end{bmatrix}, \begin{bmatrix} \mathbf{C} + \mathbf{\Sigma} & \mathbf{C}(\mathbf{x}, \mathbf{x}_*) \\ \mathbf{C}(\mathbf{x}, \mathbf{x}_*)^T & \mathbf{C}_{**} + \mathbf{\Sigma}_{**} \end{bmatrix} \right)$$

where $\mathbf{C}(\mathbf{x}, \mathbf{x}_*)$ is the covariance matrix between training inputs $\mathbf{x}$ and test inputs $\mathbf{x}_*$, $\mathbf{C}_{**}$ is the covariance matrix of $\mathbf{x}_*$ and $m_{**} = m(\mathbf{x}_*)$. The MVN formulas then imply that

$$\begin{aligned} p(\mathbf{y}_*|\mathbf{y}) \sim \mathcal{N}(\mathbf{m}_*(\mathbf{x}_*), \mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*)) \qquad & \text{where} \\ \mathbb{E}[\mathbf{y}_*|\mathcal{D}] = \mathbf{m}_*(\mathbf{x}_*) \ &= \mathbf{m} + \mathbf{C}(\mathbf{x}, \mathbf{x}_*)^T [\mathbf{C} + \mathbf{\Sigma}]^{-1} (\mathbf{y} - \mathbf{m}); \qquad (4) \\ \mathbb{V}ar(\mathbf{y}_*|\mathcal{D}) = \mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*) \ &= \mathbf{C}_{**} + \mathbf{\Sigma}_{**} - \mathbf{C}(\mathbf{x}, \mathbf{x}_*)^T [\mathbf{C} + \mathbf{\Sigma}]^{-1} \mathbf{C}(\mathbf{x}_*, \mathbf{x}). \qquad (5) \end{aligned}$$

Note that the posterior variance $\mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*)$ is equal to the prior variance $\mathbf{C}_{**} + \mathbf{\Sigma}_{**}$ minus a positive term which reflects the information gained (relative to the prior) from the training data. Furthermore, (4)-(5) are valid for *any* $\mathbf{x}_*$, i.e. both for in-sample smoothing or for out-of-sample extrapolation.

2.2 GP Hyperparameters

To implement a GP model requires specifying its hyperparameters. Note that actual inference reduces to linear-algebraic formulas in (8)-(9), and the modeling task is to capture the spatial covariance, namely the mean and kernel functions.

1. *Mean function* is often taken to be zero or a constant, $m(x) = \beta_0$. This choice is adequate for in-sample smoothing. For long-term extrapolation we wish to capture the commonly assumed longevity features, such as higher mortality at higher ages via a linear mean function: $m(x) = \beta_0 + \beta_1^{ag} x_{ag}^n$, implying that mortality rates tend to rise exponentially in Age.

2. *Covariance function* captures the correlation between mortality rate at a given Age and Year and mortality rates at other coordinates. For example, we expect the mortality for age 70 in

2010 or $x^i = (70, 2010)$, to be more correlated with $x^j = (69, 2011)$ than with $x^j = (50, 1995)$. In this paper, we employ the squared-exponential kernel:

$$C(x^i, x^j) = \mathbf{C}_{i,j} = \eta^2 \exp\Big[- \frac{(x^i_{ag} - x^j_{ag})^2}{2\theta^2_{ag}} - \frac{(x^i_{yr} - x^j_{yr})^2}{2\theta^2_{yr}} \Big]. \tag{6}$$

Above, η^2 is the process variance: when $x^i \approx x^j$, the covariance reaches its maximum value $C(x^i, x^j) \leqslant \eta^2$; when x^i and x^j are far apart, the covariance becomes very small, $C(x^i, x^j) \approx 0$. This feature of expressing the dependence structure through a spatial perspective is central to GPs and is controlled by the hyperparameters θ_{ag} and θ_{yr} in (6) that are called characteristic length-scales. The lengthscales determine how much influence an observation has on others in the Age and Year dimensions, respectively. Note that θ_{ag} —lengthscale for Age—and θ_{yr} —lengthscale for Year—are not comparable. The overall hyperparameter set for $C(\cdot, \cdot)$ is $(\theta_{ag}, \theta_{yr}, \eta^2, \sigma^2)$.

3. *Observation Likelihood.* We assume a constant observation noise $\sigma = \text{StDev}(\epsilon^i)$ $\forall i$ which is estimated via Maximum Likelihood or Markov Chain Monte Carlo along with all other hyperparameters. While this is not entirely realistic, based on the discussion in Ludkovski et al. (2018) the impact of modified observation models is minimal. A common alternative is to assume a Poisson likelihood; however it is well known that mortality data are overdispersed, so that parametrization is also mis-specified. Table 14 in the Appendix compares the estimated noise variance σ^2 in 10 European countries to their 2016 population. Law of Large Numbers would imply a linear relationship between $1/\sigma^2$ and population size. However our results clearly show that this relationship is far from linear and large countries have relatively more noisy observations.

An important aspect that influences the goodness-of-fit of a GP model is its spatial smoothness. The squared exponential covariance kernel (6) makes the mortality curves infinitely differentiable in both Age and Year dimensions (note that the GP is defined over $x \in \mathbb{R}^2_+$ and so provides a continuous interpolation of the observed data gridded by year). This will be exploited below for computing mortality improvement factors. Moreover, the lengthscales θ affect the qualitative nature of the fitted $m_*(\cdot)$. When lengthscales are too large, the fitted curves are over-smoothed and the influence of individual data points attenuates (Rasmussen and Williams 2005). As a result, there is less flexibility in $m_*(\cdot)$; to compensate, the estimated observation noise is increased and the model under-fits. In contrast, too small lengthscales indicate over-fitting of the spatial dependence, generating high-frequency oscillations in the fitted $m_*(\cdot)$ and low observation noise σ^2.

To better capture the trends in the data we fit a parametric prior mean: $m(x) = \beta_0 + \sum_{j=1}^p \beta_j h_j(x)$, where h_j's are some fixed basis functions and the β_j's are unknown coefficients. The coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ are augmented to the kernel hyperparameters and estimated simultaneously. Let $\mathbf{h}(x) = \big(h_1(x), ..., h_p(x)\big)$ and $\mathbf{H} = \big(\mathbf{h}(x^1), ..., \mathbf{h}(x^N)\big)$, then the estimation of $\boldsymbol{\beta}$ along with the predicted posterior mean and variance $s^2_*(x_*)$ for a new input x_* is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T(\mathbf{C}+\boldsymbol{\Sigma})^{-1}\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{C}+\boldsymbol{\Sigma})^{-1}\mathbf{y}; \tag{7}$$

$$m_*(x_*) = \mathbf{h}(x_*)^T\hat{\boldsymbol{\beta}} + \mathbf{c}(x_*)^T(\mathbf{C}+\boldsymbol{\Sigma})^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}); \tag{8}$$

$$\begin{aligned} s^2_*(x_*) &= C(x_*, x_*) + \\ &+ (\mathbf{h}(x_*)^T - \mathbf{c}(x_*)^T(\mathbf{C}+\boldsymbol{\Sigma})^{-1}\mathbf{H})^T(\mathbf{H}^T(\mathbf{C}+\boldsymbol{\Sigma})^{-1}\mathbf{H})^{-1}(\mathbf{h}(x_*)^T - \mathbf{c}(x_*)^T(\mathbf{C}+\boldsymbol{\Sigma})^{-1}\mathbf{H}). \end{aligned} \tag{9}$$

We note that the mean and kernel functions *interact*: choosing the mean function is analogous to de-trending, and choosing the covariance function is analogous to modeling the residuals. An informative mean function will imply that the residuals are smaller (lower η^2) and de-correlated (small θ's) compared to assuming a constant mean, which will lead to high η^2 and larger θ's.

Estimating the parameters. Our overall set of hyper-parameters is $\Theta = (\theta_{ag}, \theta_{yr}, \eta^2, \sigma^2, \boldsymbol{\beta})$. We can learn values of the hyperparameters via optimization of the marginal likelihood function which is the integral of the likelihood times the prior: $p(\mathbf{y}|\mathbf{x}, \Theta) = \int p(\mathbf{y}|\mathbf{f}, \Theta)p(\mathbf{f}|\mathbf{x}, \Theta)d\mathbf{f}$. Since $p(\mathbf{y}|\mathbf{x}, \Theta) = \mathcal{N}(\mathbf{m}, \mathbf{C} + \mathbf{\Sigma})$ and if we assume the mean function is known or fixed, the log-likelihood of the marginal is simply a MVN density:

$$\log p(\mathbf{y}|\mathbf{x}, \Theta) = -\frac{1}{2}\mathbf{y}^T(\mathbf{C} + \mathbf{\Sigma})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{C} + \mathbf{\Sigma}| - \frac{N}{2}\log(2\pi). \tag{10}$$

Thus, we have to solve a system of nonlinear equations to maximize (10) which yields the MLE estimate. We implement GP fitting via the function `km()` from the package `DiceKriging` (Roustant et al. 2012) in `R`. That package carries out MLE of Θ using a genetic optimization algorithm.

2.3 Bayesian Gaussian Process Regression

The GP hyperparameters summarize the covariance structure of the fitted mortality model. The MLE method provides a point estimate Θ_{MLE} of that structure, i.e. a "best guess" of a GP surface that fits the data. Uncertainty quantification is a major component of our analysis, in particular in assessing how similar or different are the various populations. To this end, we aim to quantify model risk, i.e. the range of GP models that are consistent with the data via a Bayesian formulation. The Bayesian GP starts with a prior on Θ and then integrates out the likelihood of the observed data to obtain the posterior distribution of the hyperparameters. A point estimate of Θ is additionally obtained from the maximum a posteriori (MAP) hyperparameters, $\Theta_{MAP} = \text{argmax}_{\Theta} \sum_i \log p(y_i|\Theta)p(\Theta)$. In fact, MLE can be viewed as a special case of MAP with improper uniform priors. In our analysis, we employ weakly informative priors to minimize influence of a priori assumptions (so that the data speaks for itself) but still regularize inference by keeping hyperparameters within reasonable ranges.

In practice, computing the posterior density $p(\Theta|\mathcal{D})$ requires to evaluate an intractable multidimensional integral. MCMC algorithms bypass this challenge by drawing samples $\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(M)}$ from the posterior. Traditionally, MCMC sampling for GP models was challenging due to strong correlation among the hyperparameters. Recently, powerful new techniques, in particular Hamiltonian Monte Carlo (HMC) have been developed to overcome this challenge. We implement Bayesian GP using `Stan` (Carpenter et al. 2017) that is built upon efficient HMC. `Stan` is a free, open-source software, written in `C++` language, and has risen to be one of the most efficient toolboxes to perform Bayesian inference and optimization for statistical models.

Following `Stan` recommendations, we standardize the input covariates (by subtracting the mean and dividing by the standard deviation, $x^i_{ag,std} := (x^i_{ag} - \mu_{\mathbf{x}_{ag}})/\sigma_{\mathbf{x}_{ag}}$) to reduce the autocorrelation between the hyperparameters and thus increase the efficiency in the MCMC chains. HMC in `Stan` further helps to cope with this autocorrelation. `Stan` returns a set of posterior MCMC samples for $\boldsymbol{\beta}$ and Θ based on standardized data, so we then have to convert these values back to the original scales. For instance, the sampled hyperparameters $\beta^{std}_{\cdot}$ of the linear mean function are transformed back by:

$$\begin{aligned} m(x^i) &= \beta_0 + \beta_1^{ag} x^i_{ag} = \beta_0 + \beta_1^{ag}(x^i_{ag,std}\sigma_{\mathbf{x}_{ag}} + \mu_{\mathbf{x}_{ag}}) \\ &= \left(\beta_0 + \beta_1^{ag}\mu_{\mathbf{x}_{ag}}\right) + \beta_1^{ag}\sigma_{\mathbf{x}_{ag}} x^i_{ag,std} \end{aligned}$$

Thus: $\beta_1^{ag} = \frac{\beta_1^{ag,std}}{\sigma_{\mathbf{x}_{ag}}}$ and $\beta_0 = \beta_0^{std} - \left(\frac{\beta_1^{ag,std}}{\sigma_{\mathbf{x}_{ag}}}\right)\mu_{\mathbf{x}_{ag}}$; in similar fashion, we can transform the lengthscales in the covariance kernel: $\theta_{ag} = \sigma_{\mathbf{x}_{ag}}\theta_{ag}^{std}$ and $\theta_{yr} = \sigma_{\mathbf{x}_{yr}}\theta_{yr}^{std}$.

For Bayesian GP hyper-priors we take $\beta_0 \sim \mathcal{N}(-4, 0.5)$, $\beta_1^{ag} \sim \mathcal{N}(0, 0.5)$. Inverse-Gamma priors are chosen for the covariance hyperparameters: $\theta_{ag}^{std} \sim \text{Inv-Gamma}(9, 12)$, $\theta_{yr}^{std} \sim \text{Inv-Gamma}(9, 12)$ which ensures that 99% of the respective prior is concentrated between 0.01 and 3.3, (Betancourt 2017). For the process variance, we take $\log \eta^2 \sim \mathcal{N}(-3, 1)$. Finally, the prior for observation noise is $\sigma^2 \sim \mathcal{N}_+(0, 0.5)$.

2.4 Further Model Outputs: Improvement Factors and Life Expectancy

A common way to interpret a mortality surface is via the (annual) mortality *improvement factors* which measure longevity changes year-over-year. In terms of the observations, the raw annual percentage mortality improvement is $1 - \frac{\exp\left(y(x_{ag};x_{yr})\right)}{\exp\left(y(x_{ag};x_{yr}-1)\right)}$. The smoothed improvement factor is obtained by replacing y's by the GP model posterior m_*'s:

$$\partial m_{back}^{GP}(x) := \left[1 - \frac{\exp(m_*(x_{ag}; x_{yr}))}{\exp(m_*(x_{ag}; x_{yr} - 1))}\right]. \tag{11}$$

Another way to visualize the output of a mortality model is via the resulting (conditional) life expectancy. To do so, we fix Age and Year and employ the estimated mortality rates across future ages. For an individual aged (x_{ag}, x_{yr}), (complete) life expectancy is calculated as:

$$\mathring{e}_{(x_{ag},x_{yr})} \approx \frac{1}{2} + e_{(x_{ag},x_{yr})} = \frac{1}{2} + \sum_{t=1}^{110-x_{ag}} S(t), \tag{12}$$

where the survival function $S(t)$ is evaluated recursively as $S(0) = 1$ and

$$S(t+1) = S(t)\big[1 - m_*(x_{ag+t}; x_{yr})\big]. \tag{13}$$

Because our training set is only up to Age 85, we use the supplied actuarial life tables for Ages 90+ to compute (13).

2.5 Interpreting the Single-Population Models

The previously discussed Table 1 illustrates fitted GP models to the 10 selected Male datasets. To summarize the main take-aways we provide the following comments about the various fitted hyperparameters:

- In the mean function, the intercept coefficient β_0 and the linear coefficient β_1^{ag} determine the shape of the mortality curve.
  - The positive β_1^{ag} of Age in Table 1 matches our expectation of an increasing Age factor. For example, $\beta_1^{ag} \approx 0.099$ in Austria implies mortality rates increase by 9.9% on average for each year of increase in Age.
  - The intercept β_0 determines the "average" baseline level of log-mortality after removing the Age effect. Populations with higher life expectancy should have lower β_0. In Table 1, Sweden and Switzerland have lowest β_0 and indeed have the highest life expectancy in this group.

- θ_{ag} controls the Age-correlation effect. For instance, $\theta_{ag} = 6.51$ in Austria can be interpreted as Ages being correlated to about $\pm 2\theta_{ag} = \pm 13$ neighboring Age-groups. This can be observed in Figure 1: populations with larger θ_{ag} have much flatter mortality improvement factors (i.e. less Age-dependent and more correlated). Namely, θ_{ag} is intuitively the "frequency" of fluctuations in $\partial m^{GP}_{back}(\cdot)$, see the 3-4 inferred "waves" in Austrian MI in the Figure over a span of 35 Age groups (50–84); while Denmark has only two such "waves" due to $\theta_{ag} = 11.30$.
- Similarly, θ_{yr} determines the influence of historical trends on the current mortality experience. A typical value of $\theta_{yr} \approx 5.78$ in Austria implies that historical patterns de-correlate and disappear after about a decade. Demographic knowledge implies that the Age structure is more persistent than the historical/temporal structure, so we expect $\theta_{ag} > \theta_{yr}$.
- Most countries have similar dependence structures where $\theta_{ag} \in [6, 14]$ and $\theta_{yr} \in [4, 10]$. In some smaller populations, such as Estonia and Lithuania, θ_{yr} is rather small, suggesting a lot of Year-over-Year variability in mortality evolution.
- The observation noise σ^2 represents the credibility of the observed raw mortality rates. Results in Table 1 and Table 14 show large countries have less noisy observations.
- Heatmaps of the fitted GP model residuals (see Appendix A) show appropriate goodness of fit without any discernible spatial patterns in either Age- or Year- dimensions. This validates the use of a spatial model and effective capturing of the underlying mortality trends.

3 Joint Modeling

In this section we proceed to set up the framework to incorporate information in mortality across different populations by pooling them into one single dataset. In particular we want to pool data from populations with similarities in mortality (e.g., countries within the same region, male and female population in a country, different states in the same country).

Data aggregation is done by treating Population as categorical input. Let L be the number of different populations considered. We now generate L factor levels, with $l = 1, \ldots, L$ the code representing each population. This is encoded as additional input dimensions, with each additional dimension coded as 1 or 0; see (Duvenaud 2014, Chu and Ghahramani 2005, Garrido-Merchán and Hernández-Lobato 2018) for further discussion of GP modeling with categorical covariates. Thus, the new input vector for the nth observation in the joint model is: $x^n = (x^n_{ag}, x^n_{yr}, x^n_{pop,1}, \ldots, x^n_{pop,L})$ where each $x^n_{pop,l}$ $(l = 1, \ldots, L)$ is an indicator function:

$$x^n_{pop,l} = \mathbb{1}_{\{\text{population}=l\}} = \begin{cases} 1 & \text{if population} = l \text{ (the } n\text{th observation is from population } l\text{);} \\ 0 & \text{if population} \neq l. \end{cases}$$

To construct a covariance kernel for the joint model, we multiply a kernel for the numerical covariates x_{ag}, x_{yr} with a kernel for the categorical ones (Qian et al. 2008, Roustant et al. 2018). Let:

$$\tilde{C}_{i,j} := \exp\left[-\frac{(x^i_{ag} - x^j_{ag})^2}{2\theta^2_{ag}} - \frac{(x^i_{yr} - x^j_{yr})^2}{2\theta^2_{yr}} \right]; \tag{14}$$

$$\Gamma_{i,j,l_1,l_2} = \exp\left[-\theta_{l_1,l_2}\delta^{ij}_{l_1,l_2} \right] \qquad \text{where} \quad l_1, l_2 \in \{1, \ldots, L\}, \tag{15}$$

with

$$\delta^{ij}_{l_1,l_2} = \begin{cases} 1 & i\text{th and } j\text{th observation come from populations } l_1 \text{ and } l_2; \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\delta^{ij}_{l_1,l_2} = 1_{\{x^i_{l_1} \neq x^j_{l_1}\}} \cdot 1_{\{x^i_{l_2} \neq x^j_{l_2}\}}$ is symmetric in i and j.

Then, the covariance between input rows x^i and x^j is set as follows:

$$\begin{aligned} C(x^i, x^j) &= \eta^2 \exp\left[- \frac{(x^i_{ag} - x^j_{ag})^2}{2\theta^2_{ag}} - \frac{(x^i_{yr} - x^j_{yr})^2}{2\theta^2_{yr}} \right] \prod_{\{l_1,l_2\}} \exp\left[- \theta_{l_1,l_2} \delta^{ij}_{l_1,l_2} \right] \\ &= \begin{cases} \eta^2 \tilde{C}_{i,j} & \text{if observations are from the same population;} \\ \eta^2 \tilde{C}_{i,j} \Gamma_{i,j,l_1,l_2} & \text{if observations from populations } l_1, l_2, \end{cases} \end{aligned} \tag{16}$$

When observations are from the same country, the covariance between the ith and jth observation is the same as in a single-population model, cf. Equation (6). Intuitively, Γ_{l_1,l_2}'s then discount the covariance when observations are from different populations, $\Gamma_{l_1,l_2} < 1$. In Equation (16), Γ_{l_1,l_2} is the function of the parameter θ_{l_1,l_2}: large value of θ_{l_1,l_2} implies low correlation between the two populations. Specifically, the correlation coefficient is $r_{l_1,l_2} := \exp\big(- \theta_{l_1,l_2} \big)$. Table 5 shows the fitted joint GP model for {Denmark, France, Sweden, UK} $\equiv \{1, 2, 3, 4\}$ Males, with population lengthscales: $\theta_{21} = 1.2602$, $\theta_{31} = 1.4569$, $\theta_{41} = 0.0945$, $\theta_{32} = 0.5123$, $\theta_{42} = 0.4869$, and $\theta_{43} = 1.2196$; assuming cells have the same Age and Year values, the cross-population correlation matrix is:

$$\begin{bmatrix} r_{21} & & \\ r_{31} & r_{32} & \\ r_{41} & r_{42} & r_{43} \end{bmatrix} = \begin{bmatrix} 0.28 & & \\ 0.23 & 0.60 & \\ 0.91 & 0.61 & 0.30 \end{bmatrix}.$$

Thus, mortality rates in UK and Denmark are highly correlated ($r_{41} = \exp(-0.0945) = 0.91$), while Sweden and France are little correlated with Denmark ($r_{21} = 0.28$ and $r_{31} = 0.23$ respectively).

Observation Noise: When modeling data from multiple populations, the observation noise variance σ^2 changes for each of the different populations. We want to maintain the homogeneity of noise variance within data from the same country and account for heterogeneous characteristics when observations from multiple populations are combined. Consequently, the variance of an observation from population l in (2) is taken to be $Var(\epsilon^i) = \sigma^2_l$ where each σ^2_l is set to the output from individual GP model for population l. This is implemented using the `noise.var` option within the `km()` call to `DiceKriging` (Roustant et al. 2012).

Mean Function: For the mean function in a multi-population model one choice is to make it the same across all populations, say $m(x^n) = \beta_0$. Alternatively, we may use a linear mean function to take into account the different trends across populations:

$$m(x^n) = \beta_0 + \beta^{ag}_1 x^n_{ag} + \sum_{l=2}^{L} \beta_{pop,l} x^n_{pop,l}. \tag{17}$$

Analogous to the coefficients of categorical covariates in regression, $\beta_{pop,l}$ can be interpreted as the mean difference between log mortality in population l and the baseline. Note that (17) implies the *same* shared prior Age structure in all populations.

Bayesian hyper-priors: When fitting GP simultaneously for multiple countries, the lengthscales $\theta_{pop,l}$ for the Population factor are added into the hyperparameter vector. Similar to the chosen priors in individual GPs, the priors in joint-population GP are: $\beta_0 \sim \mathcal{N}(-4, 0.5)$ and β_1's $\sim \mathcal{N}(0, 0.5)$; $\theta_{ag}^{std} \sim$ Inv-Gamma$(9, 12)$, $\theta_{yr}^{std} \sim$ Inv-Gamma$(9, 12)$, and $\log \eta^2 \sim \mathcal{N}(-1, 1)$. For the population lengthscales we use $\log \theta_{l_1,l_2} \sim \mathcal{N}(-1, 1)$ for all l_1, l_2. The prior for the observation noise in population l is $\sigma_l^2 \sim \mathcal{N}_+(0, 0.5)$.

3.1 Qualitative Features of Multi-population Models

Having set up the joint mortality model, let us recap its main features compared to the individual-country analysis as presented in Figure 1 and Table 1:

1. An aggregate model is expected to be more accurate, and have narrower credible bands. Recall that (95%) credible bands are $[m_*(x_*) \pm 1.96 s_*(x_*)]$, driven by the posterior standard deviation $s_*(x)$; we expect $s_*^{Joint} < s_*^{Indiv}$ due to having a larger dataset.

2. Fusing multiple populations reduces hyper-parameter uncertainty and helps to discover the "global" covariance structure. In particular, a joint model will achieve shrinkage across individual mortality improvement rates. A joint Bayesian GP model will further have tighter hyperparameter posteriors.

3. An aggregate model can be used to borrow the latest information from other country(ies) to improve prediction about the latest domestic mortality. This is especially relevant with notched datasets where there is relatively more data in other populations.

4. A joint model can achieve long-range coherence or convergence across populations, both in terms of mean forecast and individual stochastic scenarios. In particular, one may explicitly specify the long-range spread between mortality experiences.

Last but not least, a joint model is important for "slicing-and-dicing" the dataset. A challenge intrinsic to any spatio-temporal paradigm concerns the underlying assumption of covariance stationarity. Indeed, the GP model implies that the correlation structure is homogeneous across the input space. This means for instance that the correlation between Age 50 and Age 60 is the same as correlation between Age 75 and 85. Demographically, one would expect the old ages to be more correlated, hence the above assumption might not be valid. The mis-specification in turn strongly affects both the shape and width of the CI and suggests to build a model that is segmented by Age. However, for individual countries this is problematic as credibility gets lost and model inference becomes weaker. A combination of multiple countries can be used to boost credibility, and provide reliable estimates of the respective correlation structures.

4 Results

4.1 Performance metrics

To assess model performance we employ three different metrics. First, we consider out-of-sample predictive accuracy, comparing observed future mortality to its mean model forecast. The most common choice is root mean squared error (RMSE); however RMSE is highly sensitive to outliers and also to the pattern that mortality errors will be necessarily larger at higher Ages. To remedy this, we focus on the mean absolute percentage error (MAPE) measure, specifically its symmetric (SMAPE) version that corrects for the tendency of MAPE to put heavier penalties on over-estimating the observations (Armstrong and Collopy 1992, Makridakis 1993):

$$\text{SMAPE} := \frac{100}{M} \sum_{i=1}^{M} \frac{|\mathcal{Y}_*^i - m_*(x_*^i)|}{(|\mathcal{Y}_*^i| + |m_*(x_*^i)|)/2}, \tag{18}$$

where $\mathcal{Y}_*^i$ is the realized observed value at test input x_*^i and $m_*(x_*^i)$ is the predicted log-mortality rate by the model. Unlike the squared errors, SMAPE is a scale-independent measure that is convenient to compare across different data sets.

4.2 Case Study I: Two Nordic Countries

As a first illustration, we build a joint GP model, as proposed in Section 3, for Male mortality data across Sweden and Denmark ($l = 1$: Denmark and $l = 2$: Sweden). The two countries share similar demographic characteristics, such as population size and are Nordic neighbors. We test two different mean functions:

Example 1: Common constant mean function across both populations, $m(x^n) = \beta_0$.

Example 2: Linear mean function that takes into account the separation in mortality between Denmark and Sweden:

$$m(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \beta_{pop,2} x_{pop,2}^n. \tag{19}$$

Analogous to a coefficient of categorical covariates in regression, $\beta_{pop,2} \equiv \beta_{SWE}$ can be interpreted as the mean difference between log mortality in Sweden and in the baseline country, Denmark.

Table 2 shows the output of a fitted joint GP model compared with GP models fitted separately on each country. We observe that all joint hyper-parameters fall generally between those of single-population models, illustrating hyper-parameter shrinkage. In Example 2, the coefficient $\beta_{SWE} = .0794$ implies that on average male mortality in Sweden is higher than that in Denmark by 7.9%. The lengthscale $\theta_{DNK,SWE}$ controls the correlation between Denmark and Sweden, see Section 3. We notice that $\theta_{DNK,SWE}$ in Example 1 (≈ 0.0684) is smaller than in Example 2 (≈ 0.3973). This can be explained based on the choice of the mean function. When we assume both countries to share the same constant trend, it induces a higher correlation between two populations.

Table 13 in Appendix B further compares the estimated $\mathring{e}_{(x_{ag},x_{yr})}$ (complete life expectancy) in Denmark and Sweden from three different approaches: individual GP model, joint GP, and the HMD-provided life table from year 2013.

4.3 Impact of Joint Modeling on Prediction Quality

Table 3 displays out-of-sample prediction for male mortality in Sweden via the individual GP and joint (Sweden + Denmark) GP models. We recall that $m_*(x_*)$ denotes the predicted posterior mean (see Section 2.1), while $s_*^2(y_*) = C_*(x_*, x_*) + \sigma_{SWE}^2$ corresponds to the posterior marginal variance of $y(x_*)$. For prediction in the near-term future such as in 2013 and at Ages that are within or close to the training range $[70, 84]$, we see no significant difference in the prediction performance between the two approaches. We do observe that the posterior variance in the joint model is smaller than one from the individual model. This validates the advantage of a joint model to strengthen the confidence in estimating the lengthscales. One way to conveniently compare the performance in prediction between GP models is to visualize the outputs, see Figure 3. While for Denmark, the differences are very slight, in Sweden model prediction starts to diverge for $x_{yr} \geqslant 2013$ (models were fitted up to 2012). We see that the predicted curves produced by a joint model are closer to

Table 2: Comparison between single-population GP models and a joint GP model for Males in Sweden and Denmark. The training set is Ages 70–84 and Years 1990–2012.

Parameters	Denmark	Sweden	Denmark & Sweden
Constant mean: $m(x) = \beta_0$ for both countries			
β_0	-4.2190	-3.5037	-3.0996
θ_{ag}	61.3880	28.8684	36.6561
θ_{yr}	56.3974	105.7308	43.2589
$\theta_{DNK,SWE}$	–	–	0.0684
η^2	10.7882	5.1866	3.9685
σ^2_{DNK}	1.531×10^{-3}	–	1.531×10^{-3}
σ^2_{SWE}	–	8.418×10^{-4}	8.418×10^{-4}
Linear mean function $m(x)$ from 19			
β_0	-10.5627	-11.2166	-11.0174
β_1^{ag}	0.0984	0.1086	0.1046
β_{SWE}	–	–	0.0794
θ_{ag}	30.9987	19.4562	22.4098
θ_{yr}	19.5434	10.7198	23.2661
$\theta_{DNK,SWE}$	–	–	0.3973
η^2	0.1223	0.0416	0.1030
σ^2_{DNK}	1.516×10^{-3}	–	1.516×10^{-3}
σ^2_{SWE}	–	8.025×10^{-4}	8.025×10^{-4}

Table 3: Predicted Male Swedish mortality in 2013 and 2016 using a joint Denmark + Sweden GP model as in Table 2.

	Year 2013				Observed	Year 2016				Observed
	Single		Joint GP		value μ	Single		Joint GP		value μ
Age	m_*	$s_*(y)$	m_*	$s_*(y)$	(2013)	m_*	$s_*(y)$	m_*	$s_*(y)$	(2016)
75	-3.4501	0.0300	-3.4682	0.0292	-3.4526	-3.4542	0.0399	-3.5285	0.0316	-3.5518
85	-2.2000	0.0330	-2.2189	0.0309	-2.2462	-2.1996	0.0461	-2.2578	0.0345	-2.2835
90	-1.5615	0.0473	-1.5761	0.0406	-1.6602	-1.5641	0.0637	-1.6029	0.0461	-1.6575

the observed values in the test period from 2013-2016. Table 4 compares the predictive accuracy between the models via SMAPE and confirms that joint GP is better (smaller SMAPE values) at

Table 4: Prediction accuracy via SMAPE for single-population and joint GP models from Table 2. The test set is Ages 70–84 in Years 2013, 2015, and 2016.

SMAPE		2013 (one-year out)		2015 (three-year out)		2016 (four-year out)	
		Single	Joint GP	Single	Joint GP	Single	Joint GP
Age $\in [70, 84]$	Denmark	1.5798	**1.4451**	1.3445	**1.2862**	1.2584	**1.1955**
	Sweden	1.0450	**0.8256**	1.9752	**1.1011**	2.5272	**0.9038**

Table 5: Joint model using mortality rates in 4 countries: Denmark, France, Sweden, and UK. The aggregated training dataset contains Ages 70–84 and Years 1990–2012 for Males.

Mean function			Covariance hyper-parameters						
β_0	β_1^{ag}	β_{SWE}	θ_{ag}	θ_{yr}	$\theta_{DNK,FRA}$	$\theta_{DNK,SWE}$	$\theta_{DNK,UK}$	$\theta_{FRA,SWE}$	$\theta_{FRA,UK}$
-10.5417	0.1006	-0.0268	14.0396	9.5543	1.2602	1.4569	0.0945	0.5123	0.4869
–	β_{FRA}	β_{UK}	–	$\theta_{SWE,UK}$	η^2	σ^2_{DEN}	σ^2_{FRA}	σ^2_{SWE}	σ^2_{UK}
–	-0.0869	0.0069	–	0.7036	0.0327	1.516×10^{-3}	3.393×10^{-4}	8.021×10^{-4}	6.887×10^{-4}

out-of-sample prediction in both populations.

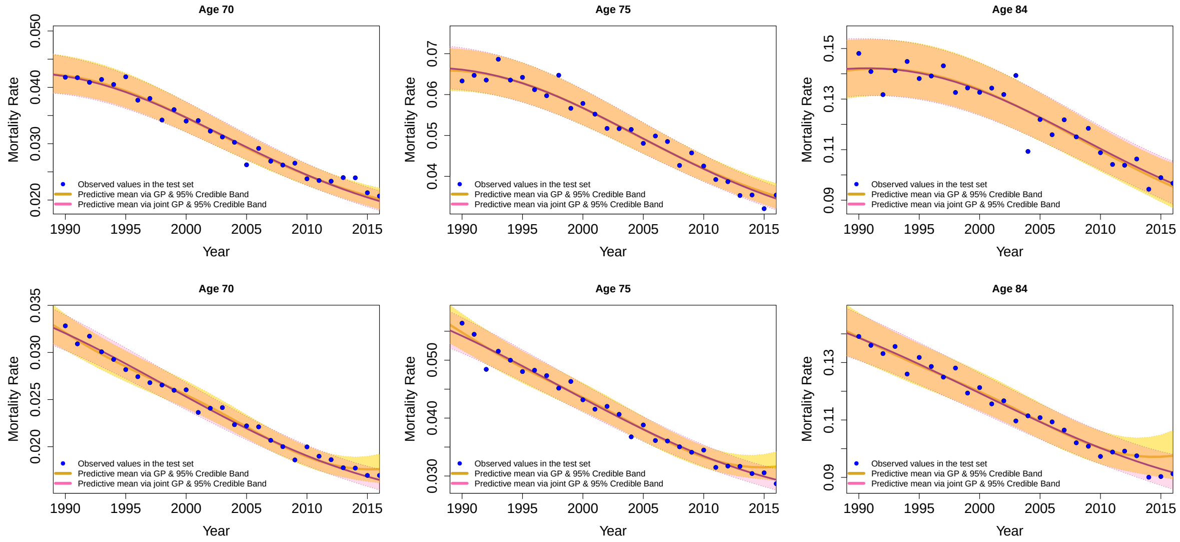
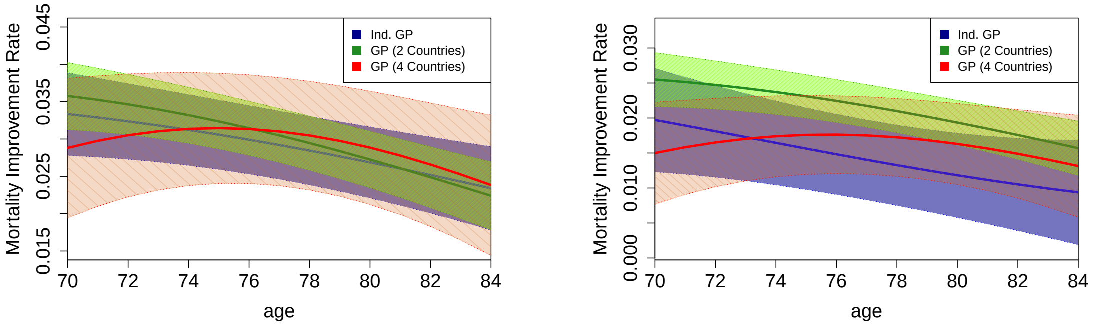


Figure 3: 95% credible intervals for observed log-mortality $y(x_*)$ across the individual and joint GP models. Top row: Denmark Males; bottom row: Sweden Males. Note that for up to 2011, the smoothed mortality curves and CIs are essentially identical for both approaches.

4.4 Case Study II: Four European Nations

We can straightforwardly implement the joint GP framework to model the mortality for more than 2 populations. Table 5 demonstrates a joint GP model on four countries: Denmark, France, Sweden, and UK, trained on Males aged 70–84 and Years 1990–2012. In Equation (17), the coefficient β_1^{ag} provides the log-linear Age pattern across all populations, while the differences in baseline mortality (Denmark) are captured through the coefficients β_{pop}'s. (We will show that these differ-



(a) Denmark Males

(b) Sweden Males

Figure 4: Comparison of annual mortality improvement factors between different joint models. Besides the mean of improvement factors $\partial m^{GP}_{back}(ag; 2012)$ (11) for Ages $70, \ldots, 84$, we also show the respective 95% posterior credible band.

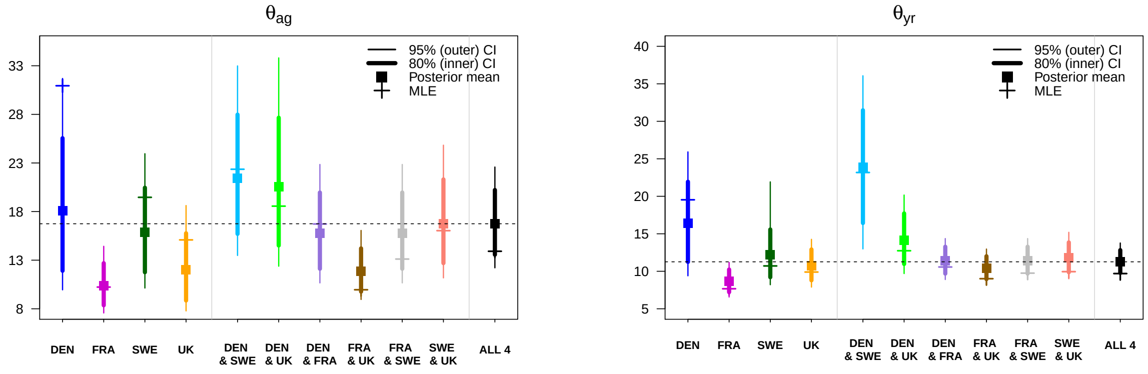
ences are actually insignificant after we fit a fully Bayesian GP on this dataset.) In the covariance kernel, the θ_{ag} and θ_{yr} are shared between all populations, while θ_{l_1,l_2}'s control the cross-population correlations.

Figure 4 examines the predicted annual mortality improvement factors between individual and different joint models, concentrating on Denmark and Sweden. Large θ_{ag} lengthscales in individual GP models lead to essentially linear improvement rate factors (blue curves). When modeling Sweden and Denmark together (green curves), lengthscale decreases and $s_*(x_*)$ falls, so the improvement rate factors become more Age-dependent and with tighter credible bands. This effect becomes even stronger when we use all four populations together. The corresponding smoothed curves (colored in red) are quite nonlinear, and in particular imply that improvement at young Ages (< 60) has slowed dramatically. This illustrates that a joint model is better able to distinguish between signal and "noise" and therefore pick up divergent changes in mortality faster, while a single model would often smooth latest changes away.

4.5 Hyperparameters in Joint Models

We perform Bayesian GP on the above 4 populations: Sweden, Denmark, France, and UK. Comparison between the resulting maximum likelihood (MLE) and maximum a posteriori probability (MAP) estimates is shown in Table 6. The MLE fits fall within the 95% posterior credible intervals from the `Stan` model. The 95% credible interval for β_1^{ag} confirms the significance of the linear effect of Age. The mean function coefficients $\beta_{pop,l}$'s in the joint model estimate the mean differences in mortality between Sweden (the baseline) and other countries. The 95% posterior CI's for these coefficients all contain 0, implying that they are not statistically significant. This indicates that there is no clear difference in the respective mortality experience which is intuitive since all populations are from developed countries within the same geographic area.

Figure 5 shows the inferences of the lengthscales for Age and Year along with MLE estimations when fitting mortality separately for each country: Denmark, France, Sweden, and UK, versus jointly modeling them as groups of 2, or jointly as all 4 together. The figures visualize how joint GP models produce tighter hyperparameter posteriors. For example, the posterior mean of θ_{ag} in Denmark is relatively large and its credible bands are wide compared to the other three countries



(a) Length-scale in the Age dimension

(b) Length-scale in the Year dimension

Figure 5: `Stan` MCMC posteriors of the lengthscales θ for Age and Year across populations and joint models with different groupings. The +'s indicate the respective MLE estimates from a `DiceKriging` model. The dashed lines indicate the MCMC MAP estimate from the 4-population joint model.

(Figure 5.a). However, once we pair Denmark with either Sweden, UK, or France (Figure 5.a —light blue, light green, and purple CIs respectively), the credible bands of θ_{ag} become narrower and in the more reasonable range of $\theta_{ag} \in [15, 30]$. This effect is even further amplified when taking all 4 countries together. The underlying concept is that the more populations are added into the model, the closer we get at discovering the "universal" representation of mortality pattern. In Figure 5, the 4-population MAP estimates of the lengthscales (dashed horizontal lines) intersect with a majority of CIs suggesting that there is indeed a common covariance structure which is gradually revealed as we increase the training dataset.

The posterior MCMC samples of the lengthscales θ_{l_1,l_2}'s can be used to sample the posterior distribution of the correlation in mortality rates between a pair of countries in the model. In this spirit, Table 7 reports the posterior mean (bold numbers) and the respective 95% credible bands of R_{l_1,l_2}. We note that the credible bands are quite wide, so the model is not too confident about cross-population correlations.

4.6 Case Study III: Joint Modeling of Male and Female Datasets

The gender gap in mortality varies by country but Males tend to have higher mortality rates than Females due to both biological and non-biological factors (Hazzard 1986, Kraemer 2000, Regan and Partridge 2013). For example, women outlive men by 7 years on average in developed countries (United Nations 2011). Modeling mortality for each gender separately often fails to take into account the interdependent relationship between them and further results in divergent and implausible long-run forecasts even if the same fitting procedure is applied. In this section we demonstrate how a joint GP framework can be used to simultaneously model Male and Female mortality within a country, using Denmark as the case study. Treating Female as a baseline factor, the coefficient $\beta_M \approx .4157$ in the fitted GP mean function reveals the higher mortality (by 40% on average) for males compared to females. As we will see below, β_M plays an important role in achieving coherent long-term forecasts as we expect Females to continue having lower mortality than Males. The length-scale $\theta_{F,M} \approx 0.4925$ confirms the correlation between mortality of the two genders.

Table 6: Hyper-parameter estimates based on maximum likelihood (`DiceKriging`) and maximum a posteriori probability (`Stan`), along with MCMC summary statistics using a joint mortality model across four countries: Denmark, Sweden, France, and UK. Training set contains Males aged 70–84 during Years 1990–2012. Sweden used as baseline population.

Parameters	`DiceKriging`	`Stan`		
	MLE	MAP	MCMC Mean	MCMC 95% Posterior CI
β_0	-10.5417	-10.0220	-10.5337	$(-12.0847, -9.1274)$
β_1^{ag}	0.1006	0.0958	0.0967	(0.0847, 0.1085)
β_{SWE}	-0.0268	-0.0685	0.1239	$(-0.2438, 0.5827)$
β_{FRA}	-0.0869	-0.0971	-0.0060	$(-0.3596, 0.3844)$
β_{UK}	0.0069	0.000	0.1122	$(-0.2252, 0.4961)$
θ_{ag}	14.0396	12.1915	17.4166	(12.0294, 24.0641)
θ_{yr}	9.5543	9.2694	11.3858	(8.2536, 13.3009)
$\theta_{DNK,FRA}$	1.2602	0.3773	0.8269	(0.1544, 2.9089)
$\theta_{DNK,SWE}$	1.4569	0.2725	0.5094	(0.0889, 1.8891)
$\theta_{DNK,UK}$	0.0945	0.0799	0.1579	(0.0286, 0.5473)
$\theta_{FRA,SWE}$	0.5123	0.1943	0.3658	(0.0797, 1.0949)
$\theta_{FRA,UK}$	0.4869	0.1445	0.1439	(0.0383, 0.3917)
$\theta_{SWE,UK}$	0.7036	0.1801	0.6132	(0.0530, 2.6660)
η^2	0.0327	0.0392	0.0684	(0.0289, 0.1520)
σ^2_{DEN}	1.516×10^{-3}	1.514×10^{-3}	1.528×10^{-3}	$(1.315 \times 10^{-3}, 1.772 \times 10^{-3})$
σ^2_{FRA}	3.394×10^{-4}	3.371×10^{-4}	3.459×10^{-4}	$(2.956 \times 10^{-4}, 4.045 \times 10^{-4})$
σ^2_{SWE}	8.022×10^{-4}	8.007×10^{-4}	8.226×10^{-4}	$(7.033 \times 10^{-4}, 9.640 \times 10^{-4})$
σ^2_{UK}	6.887×10^{-4}	6.849×10^{-4}	7.001×10^{-4}	$(5.985 \times 10^{-4}, 8.165 \times 10^{-4})$

Table 7: MCMC for the inferred correlation between populations in the Male Sweden-Denmark-France-UK joint model.

	SWE	DEN	FRA	UK
DEN	**0.6561**	**1**		
	(0.1512, 0.9149)	-		
FRA	**0.7198**	**0.5178**	**1**	
	(0.3345, 0.9233)	(0.0545, 0.8568)	-	
UK	**0.6340**	**0.8622**	**0.8697**	**1**
	(0.069, 0.9483)	(0.5784, 0.9712)	(0.6758, 0.9623)	-

Another effect we observe is the shrinkage in terms of the lengthscales. Table 8 reports the Age and Year lengthscales across individual and joint models. We note that the joint parameters ($\theta_{ag} \approx 12.0668$ and $\theta_{yr} \approx 9.9789$) are closer to the ones in the Female model ($\theta_{ag} \approx 11.5567$ and $\theta_{yr} \approx 9.5256$) than the Male model ($\theta_{ag} \approx 30.9566$ and $\theta_{yr} \approx 19.5687$). Thus we witness the strong effect of borrowing cross-population information in this example. Table 9 reports SMAPE on out-of-sample forecasts and confirms that joint GP performs very well even at high ages with zero information presented in the training set.

Table 8: Joint model on Ages 70–84 and Years 1990–2012 for Males and Females in Denmark.

Parameters	Female	Male	Female & Male
β_0	-11.6755	-10.5628	-11.3647
β_1^{ag}	0.1095	0.0984	0.1054
β_M	–	–	0.4157
θ_{ag}	11.5567	30.9566	12.0668
θ_{yr}	9.5256	19.5687	9.9789
$\theta_{F,M}$	–	–	0.4925
η^2	0.0429	0.1224	0.0379
σ_F^2	1.489×10^{-3}	–	1.489×10^{-3}
σ_M^2	–	1.516×10^{-3}	1.516×10^{-3}

Table 9: Prediction accuracy via SMAPE between individual- and joint-gender GP models. Training set is Ages 70–84 and Years 1990–2012 in Denmark. Test set is Ages 70–84 and Years 2013, 2015, 2016.

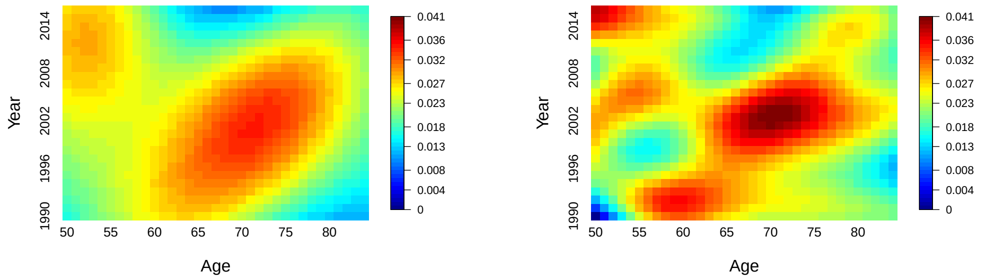
SMAPE		2013 (one-year out)		2015 (three-year out)		2016 (four-year out)	
		Single	Joint GP	Single	Joint GP	Single	Joint GP
Age $\in [70, 84]$	Female	0.9422	**0.8834**	1.8973	**1.7845**	1.4010	**1.2269**
	Male	1.5802	**1.5062**	1.3444	**1.2454**	1.2583	**1.1819**

5 Features of Joint GP Models

5.1 Improved Hyperparameter Estimation

In Table 1, the estimated length-scales in the Year dimension, θ_{yr}, are relatively large for Sweden and Switzerland (bold numbers) and relatively small for Estonia and Lithuania (italic numbers). These empirical features confirm the opportunity to *better estimate the hyperparameters* by utilizing multiple data sets. We also observe that some of the "outlier" data sets, such as Switzerland, might be resistant to accurate modeling and would strongly benefit from a more structured way of learning their correlation structure. It is known that GPs might have difficulties in estimating lengthscales, for example due to the likelihood function (10) being highly multi-modal, or conversely very flat around its maxima. Providing more data is one remedy.

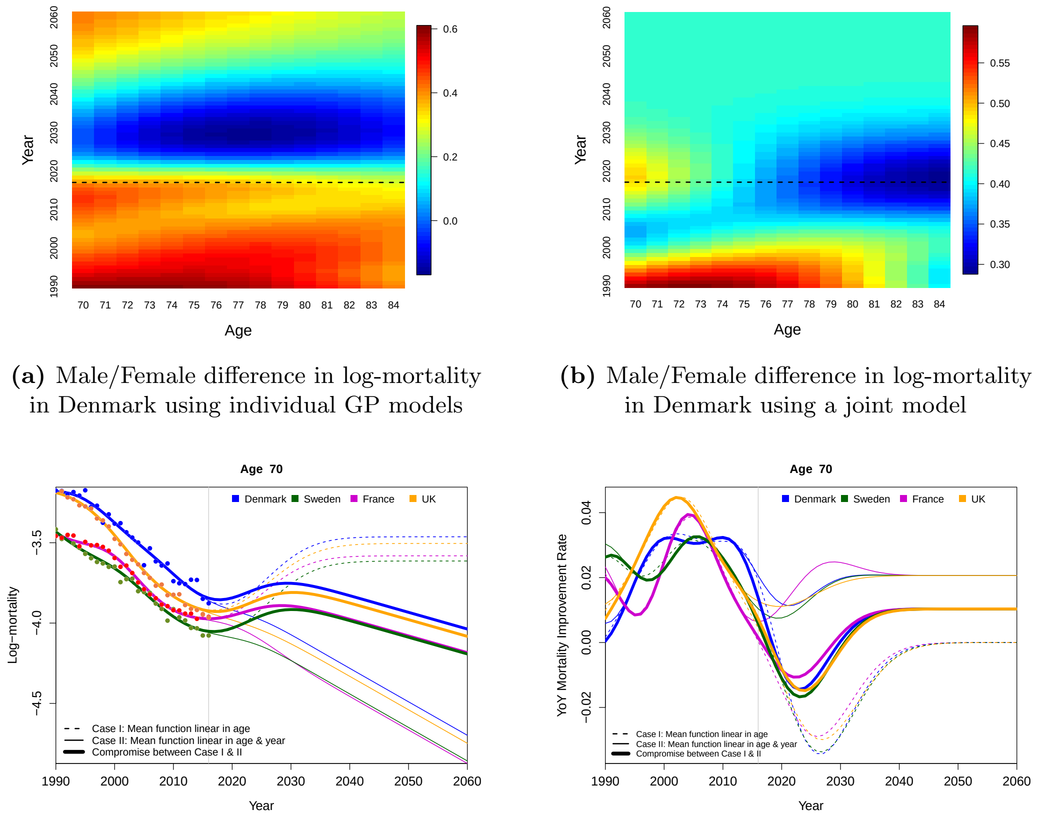
Figure 6 visualizes the fitted Swiss Male mortality improvement factors from a single-population GP model (left panel) and a joint Switzerland-Austria GP model (right panel). We observe that the individual model over-smoothes the data and hides most of the fluctuations. In contrast, pooling data across the two populations shrinks unreasonably large lengthscales and provides a much better fit in Figure 6.b.



(a) Improvement factors via individual GP

(b) Improvement factors from a joint model

Figure 6: Predicted annual Male mortality improvement factors $\partial m^{GP}_{back}(.; yr)$ in Switzerland. Training set is Ages 50–84 and Years 1990–2012.



(a) Male/Female difference in log-mortality in Denmark using individual GP models

(b) Male/Female difference in log-mortality in Denmark using a joint model

(c) Predicted Log-mortality rate for Age 70 Males using a joint GP model for Sweden, Denmark, France, and UK

(d) Annual mortality improvement factors at Age 70 Males using a joint GP model for the respective four countries.

Figure 7: Long-term mortality forecasting over years 1990 to 2060. All models are trained using Ages 70–84 and Years 1990–2016 (edge of training set indicated by the dashed lines). Top panels show the forecasted mean difference between Danish Male and Female mortality.

5.2 Coherent Mortality Forecasts

Fitting GP models for individual populations tends to generate divergent long-term forecasts that are inconsistent with historical patterns. To illustrate this issue, we fit individual GP models for Males and Females in Denmark for Ages 70–84 and calendar Years from 1990 to 2016 and then forecast gender-specific mortality forward up to 2060 (45 years into the future). The heatmap in Figure 7.a displays the resulting projected Male-Female differences in log-mortality. We observe that the models imply that as early as 2030, Males will have lower mortality than females. For example, at Age 80 Males had 52% higher mortality in 1990, 34% higher in 2016 but will have 14% *lower* mortality in 2030. This unlikely forecast is not caused by any specific feature of GP modeling, but arises “randomly” due to independent treatment of the two datasets. We note that divergence manifests itself both through implausible difference in mean forecasts, as well as excessively fast changes in relative mortality, see the rapid overtaking between the two genders in Figure 7.a.

On the contrary, forecasts based on joint models maintain the historical characteristics observed in the data into the future. Figure 7.b shows the Male-Female relative log-mortality coming from a joint GP model. In that case, the relative forecast is coherent: Females are projected to maintain higher longevity and historical patterns slow dissipate over time to the long-term spread of about $\beta_M = 41.5\%$ between same-age Male and Female mortality, cf. Table 8. The respective stochastic scenarios similarly capture the long-run dependence between the two genders.

In GP models, the long-term forecast is driven by the prior of f, and specifically by the mean function $m(\cdot)$. Thus, the relative differences in mortality between populations are controlled through the choice of $m(\cdot)$, so that different ways of achieving coherence are transparent to the modeler. To highlight this aspect, Figures 7.c and7.d show the log-mortality and annual mortality improvement rates for Males aged 70 across Sweden, Denmark, France, and UK, in the period from 1990 to 2060, estimated via a joint 4-population GP model (each country curves are shaded with a unique color). In the Figures we illustrate three different scenarios about long-term coherence:

1. Zero long-term mortality improvement, captured by the linear mean function $m(x^n) = \beta_0 + \beta_1^{ag} x^n_{ag} + \sum_{l=2}^{L} \beta_{pop,l}\big(x^n_{pop,l}\big)$ (dashed curves). All mortality improvement factors converge to zero (right panel) and the long-run mortality differences are summarized by the $\beta_{pop,l}$ coefficients.

2. Long-term mortality improvement based on historical pattern (thin solid curves). This is encapsulated via $m(x^n) = \beta_0 + \beta_1^{ag} x^n_{ag} + \beta_1^{yr} x^n_{yr} + \sum_{l=2}^{L} \beta_{pop,l} x^n_{pop,l}$. In the long-run $\partial m^{GP}_{back}(.; yr) \to \beta_1^{yr}$ (about 2% annual); again $\beta_{pop,l}$ determine the long-run relative difference in longevity of different populations.

3. Long-term mortality improvement based on expert judgement (thick solid lines). We again use $m(x^n) = \beta_0 + \beta_1^{ag} x^n_{ag} + \beta_1^{yr} x^n_{yr} + \sum_{l=2}^{L} \beta_{pop,l}\big(x^n_{pop,l}\big)$, but this time β_1^{yr} coefficient is picked by the modeler and fixed at 1%. Since it is not possible to fully extrapolate the future longevity trends from past data, it is appropriate to use expert opinions about future mortality (Booth and Tickle 2008). By way of illustration, we assume in Figures 7.c-d that future improvements will moderate to only 1% annually, reflecting recent slowdown in MI.

In all three scenarios above, we observe that the choice of $m(\cdot)$ has minimal impact on in-sample forecasts that are largely driven by the training data covering 1990–2016. On the other hand, the long-term levels of mortality improvement are *completely* driven by $m(\cdot)$. Finally, for short-term extrapolation (roughly 2016–2025 in the Figure, reflecting the fitted Year lengthscale $\theta_{yr} \simeq 10$, cf. Table 6) the forecasts blend information from the training set and from $m(\cdot)$. Note that in this example some of the individual mortality curves may cross, i.e. the relative order of longevity in different populations may change over time (such as France surpassing Sweden's longevity) reflecting relatively higher recent improvement rates. Nevertheless, we see a very strong coherence so that mortality rates across populations all move roughly in unison over time, matching our intuition about the persistent commonality of their future mortality experiences.

5.3 Incorporating Latest Data from Other Populations

In HMD, the reported data from different countries arrives non-synchronously. Indeed, the last available year of data varies from one country to another. The prevailing approach is to consider the time period that is common to all countries that are being modeled. This implies that the most recent observations may be dropped for some countries. Of course, such recent data is in fact the most informative for picking up new insights about the present longevity trend. Note that the HMD datasets are updated continuously, so that which datasets have the latest observations changes dynamically over time.

	Sweden	Denmark
Model 1	1990 - 2016	–
Model 2	1990 - 2015	–
Model 3	1990 - 2015	1990 - 2015
Model 4	1990 - 2015	1990 - 2016

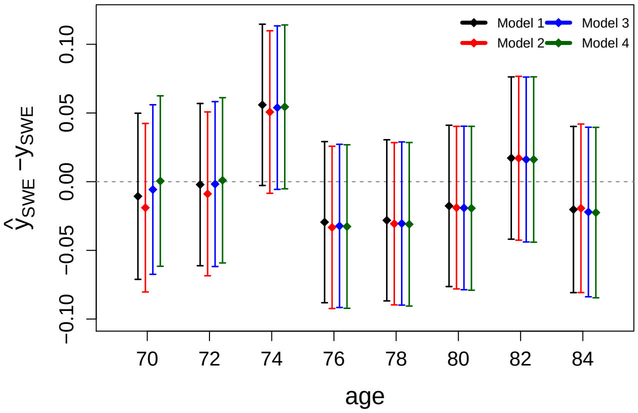


Figure 8 & Table 10: Accuracy of Male mortality predictions for Sweden at Ages $\in [70, 84]$ for calendar Year 2016 between Models 1–4. We view Model 1 as the benchmark. The right panel shows the predictive distributions of $m_*(ag; 2016)$ for the respective Ages relative to the realized 2016 observation, so that closer to zero (dashed line) is always better.

GP models can easily handle such "notched" joint datasets, allowing to fully incorporate all latest data without removing any observations. Recall that since GPs operate with a tabular representation of (x^i, y^i), one simply adds rows to incorporate more observations. In Figure 8, we illustrate the prediction of Male mortality in Sweden for the year 2016 based on several individual Swedish and joint Denmark-Sweden datasets. Our benchmark is Model 1 that already has access to 2016 Swedish data. For the remainder of this example we then assume that this data is still unavailable, so that only 2015 data is provided for Sweden, however 2016 mortality experience has already been released for Denmark. Thus, to forecast 2016 Swedish mortality one must perform a 1-year-out extrapolation. A basic choice is a single-population Model 2 that uses Swedish data up to 2015. An improvement is the joint Model 3 that uses both Danish and Swedish data up to 2015 and would be the typical way to cross-sectionally fuse mortality data. As expected (cf. Tables 3 and 4), the joint Model 3 achieves lower prediction errors relative to Model 1. Finally, Model 4 works with a notched most-up-to-date dataset that contains Danish 2016 data. Model 4 is not possible in the Lee-Carter framework that requires rectangular datasets. We observe that Model 4 materially improves on

Table 11: Prediction accuracy via SMAPE for single-population and joint models for Males in Germany and Switzerland. Training is based on the specified Age groups and Years 1990–2012. More accurate models are bolded.

SMAPE		2013 (one-year out)		2015 (three-year out)		2016 (four-year out)	
		Single*	Joint GP†	Single*	Joint GP†	Single*	Joint GP†
Age $\in [70, 84]$	Germany	0.8486	**0.7437**	1.2586	**1.0292**	1.9549	**1.4930**
	Switzerland	**1.0114**	1.1870	1.3453	**0.8290**	4.8664	**1.8771**
		Single*	Joint GP‡	Single*	Joint GP‡	Single*	Joint GP‡
Age $\in [85, 100]$	Germany	5.6969	**4.8998**	4.7352	**3.5757**	**2.8355**	5.5830
	Switzerland	5.1769	**4.4189**	5.4892	**4.4078**	10.9315	**7.6852**

$\star$: Ages 70-100 (Full), †: Ages 70-84 (Old), and ‡: Ages 85-100 (Advanced Old).

Model 3 and is practically as good as the benchmark Model 1. In other words, borrowing latest information from a neighboring population is nearly as good as having the latest domestic data, and is significantly better than just using the available domestic data (Model 2).

5.4 Age-Segmented Models

A key assumption of GP modeling is stationary covariance structure, i.e. that the covariance between input cells is fully specified by their relative distance (expressed through the lengthscales θ's.) rather than absolute coordinates. Such a stationarity assumption may be violated in practice, in particular when considering extreme ages where there may be stronger spatial correlation. A natural way to handle such model mis-specification is to build an Age-segmented model. For a single-population, segmentation is problematic as it reduces the training set size. In contrast a joint multi-population model is well-suited to such "slicing-and-dicing". In this section we investigate these aspects by building Age-segmented joint models, namely for Ages 70–84 (similar to previous case studies) and for Ages 85–100 (extreme Old). Comparing them also provides a test for spatial homogeneity. We furthermore compare to single-population models fitted on the full range of Ages 70–100 (considering only 15 Age groups in an individual model is not recommended as the training dataset is very small. This is especially so for ages 85++ where the data is very noisy.).

Table 11 compares the performance of single GP models fitted on full-Age-range German and Swiss Male mortality datasets, as well as joint 2-population GP models fitted on the 70–84 and 85–100 age segments. Throughout we use a training period of 1990–2012 and test period of 2013–2016. We purposely choose the same size of the training dataset across the models (i.e. individual models have 30 Age groups, while joint models have 15 Age groups from each of two populations) to allow a fair comparison and hence isolate the effect of Age segmentation. Table 11 shows that joint models outperform, suggesting that it is beneficial to combine data from different populations but same Age groups for prediction purposes. Note that the gains from joint modeling are largest at extreme ages where prediction errors are necessarily higher due to larger observation variance (as number of Exposed is very low).

Table 12 reports the respective fitted GP hyperparameters. We observe that while most hyperparameters are similar across the two age groups, the correlation in the Age dimension as captured by θ_{ag} changes significantly. At Ages 70-84 we obtain $\theta_{ag} \simeq 16$, in line with other models in Section 4, see Tables 2 and 5. However at Ages 85–100 we obtain $\theta_{ag} \simeq 6.8$ which implies more idiosyncratic Age effects at extreme old ages. In contrast, we find that the correlation between the two popula-

Table 12: Hyperparameters of single-population and joint GP models for German and Swiss Males in Table 11.

Parameters	Germany	Switzerland	Joint	Joint
(Ages)	(70–100)	(70–100)	(70–84)	(85–100)
β_0	-11.2886	-10.6350	-10.3927	-10.8608
β_1^{ag}	0.1100	0.0998	0.1003	0.1054
β_{SWI}	–	–	-0.1913	-0.1540
θ_{ag}	3.7445	13.3372	16.0992	6.8645
θ_{yr}	23.1392	4.4980	13.0569	11.7254
$\theta_{GER,SWI}$	–	–	0.0943	0.0371
η^2	0.1128	0.0373	0.0537	0.0497
σ^2_{GER}	1.687×10^{-3}	–	1.328×10^{-3}	1.328×10^{-3}
σ^2_{SWI}	–	4.383×10^{-3}	1.373×10^{-3}	1.373×10^{-3}

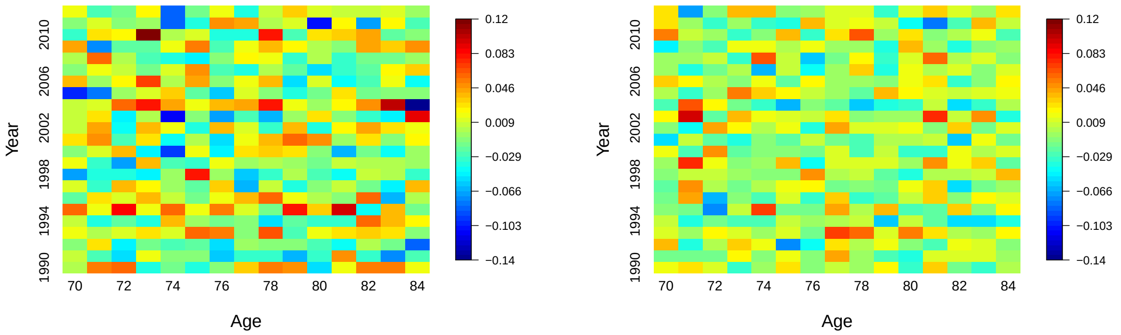
tions ($\theta_{GER,SWI}$ hyperparameter) is stronger at extreme Ages, suggesting that respective mortality improvement at ages 85++ is driven by cross-national European, rather than domestic trends.

6 Conclusion

We have investigated stochastic multi-population mortality models based on Gaussian process regression. In our approach, cross-population dependence is captured via spatial correlation based on the inferred $\theta_{pop,l}$ hyperparameters. We show that a joint model is able to deliver multiple modeling benefits, from better hyperparameter estimation to coherent joint long-term forecasts, to full fusion of the most recent neighbor mortality observations. Looking ahead, it would be worthwhile to investigate large-scale models, e.g. based on the full HMD database of 40 countries and 2 genders. This requires additional modeling infrastructure as the presented approach becomes computationally expensive for $L > 5$ populations (more than $N \gg 2500$ total cells). Fortunately, there is a very active ongoing progress on large-scale GP models especially for gridded data like in HMD, see e.g. Flaxman et al..

A different avenue of future research would be to systematically explore the best spatial covariance structures, as encapsulated by the kernel function $C(x, x')$. In this paper we focused on only using the squared-exponential kernel and standard Age- and Year-effects. It is feasible to consider further dependence formats, e.g. birth Cohort effect, and other kernel families, such as the Matern Ludkovski et al. (2018). A third direction would be to revisit the observation variance assumption via GLM (generalized linear model) GP formulations.

A Independence Assumption via Residuals



(a) Residuals from joint GP model for Denmark **(b)** Residuals from joint GP model for Sweden

Figure 9: Testing for dependence in residuals from a joint GP model.

B Impact of Joint Models on Life Expectancy

Table 13: Predicted complete life expectancy, $\mathring{e}_{(x_{ag};x_{yr})}$ for Males aged 70 and 84, from individual and joint GP models. LT refers to the HMD life table used as benchmark.

(Age,Year)	Denmark			Sweden		
	GP	Joint GP	LT	GP	Joint GP	LT
$(70, 2013)$	13.74	13.72	13.96	14.38	14.50	14.8
$(84, 2013)$	5.73	5.68	5.86	5.65	5.74	6

C Estimated Observation Noise vs Population Size

Table 14: Fitted observation noise variance in individual GP models versus population by country in 2016. Estonia is the baseline with population of 1.3 million and fitted $\sigma^2 \approx 1/155.3625$. Source: `https://ec.europa.eu/eurostat`.

	Pop'n ratio	Inverse of σ^2 ratio
Estonia	1	1
Lithuania	2.1950	2.1922
Denmark	4.3370	2.3776
Switzerland	6.3279	2.4729
Austria	6.6116	3.1352
Sweden	7.4859	3.2784
Netherlands	12.9026	5.2078
UK	49.6849	11.8718
France	50.7092	15.8156
Germany	62.4462	7.8849

References

Katrien Antonio, Sander Devriendt, Wouter de Boer, Robert de Vries, Anja De Waegenaere, Hok-Kwan Kan, Egbert Kromme, Wilbert Ouburg, Tim Schulteis, Erica Slagter, et al. Producing the Dutch and Belgian mortality projections: a stochastic multi-population standard. *European Actuarial Journal*, 7 (2):297–336, 2017.

J.Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1):69 – 80, 1992.

Michael Betancourt. Robust gaussian processes in stan, 2017. URL `https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html`.

Carl Boe, Celeste Winant, Timothy Riffe, Magali Barbieri, John R Wilmoth, Domantas Jasilionis, Pavel Grigoriev, Dmitri Jdanov, Vladimir M Shkolnikov, and Dana Glei. Data Resource Profile: The Human Mortality Database (HMD). *International Journal of Epidemiology*, 44(5):1549–1556, 2015.

Tim J Boonen and Hong Li. Modeling and forecasting mortality with economic growth: a multipopulation approach. *Demography*, 54(5):1921–1946, 2017.

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

Bob Carpenter, Andrew Gelman, Matthew D.Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Patricia Carracedo, Ana Debón, Adina Iftimi, and Francisco Montes. Detecting spatio-temporal mortality clusters of European countries by sex and age. *International journal for equity in health*, 17(1):38, 2018.

Hua Chen, Richard MacMinn, and Tao Sun. Multi-population mortality models: A factor copula approach. *Insurance: Mathematics and Economics*, 63:135–146, 2015.

Marcus C Christiansen, Evgeny Spodarev, and Verena Unseld. Differences in European mortality rates: A geometric approach on the age–period plane. *ASTIN Bulletin: The Journal of the IAA*, 45(3):477–502, 2015.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *J. Mach. Learn. Res.*, 6: 1019–1041, 2005.

Valeria D'Amato, Steven Haberman, Gabriella Piscopo, Maria Russolillo, and Lorenzo Trapani. Multiple mortality modeling in Poisson Lee–Carter framework. *Communications in Statistics-Theory and Methods*, 45(6):1723–1732, 2016.

Ana Debón, Francisco Martínez-Ruiz, and Francisco Montes. A geostatistical approach for dynamic life tables: The effect of mortality on remaining lifetime and annuities. *Insurance: Mathematics and Economics*, 47(3):327–336, 2010.

Antoine Delwarde, Michel Denuit, Montserrat Guillén, and Antoni Vidiella-i Anguera. Application of the Poisson log-bilinear projection model to the G5 mortality experience. *Belgian Actuarial Bulletin*, 6(1): 54–68, 2006.

David Kristjanson Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

Vasil Enchev, Torsten Kleinow, and Andrew JG Cairns. Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342, 2017.

Seth Flaxman, Andrew Gelman, Daniel Neill, Alex Smola, Aki Vehtari, and Andrew Gordon Wilson. Fast hierarchical Gaussian processes. Technical report, Preprint at http://sethrf.com/files/fast-hierarchical-GPs.pdf.

Eduardo C. Garrido-Merchán and Daniel Hernández-Lobato. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *CoRR*, abs/1805.03463, 2018.

Quentin Guibert, Olivier Lopez, and Pierrick Piette. Forecasting mortality rate improvements with a high-dimensional VAR. 2017.

William R. Hazzard. Biological basis of the sex differential in longevity. *Journal of the American Geriatrics Society*, 34(6):455–471, 1986.

HMD. The Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)., 2018. URL `www.mortality.org`.

Rob J Hyndman, Heather Booth, and Farah Yasmeen. Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, 50(1):261–283, 2013.

Torsten Kleinow. A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, 63:147–152, 2015.

Torsten Kleinow and Andrew JG Cairns. Mortality and smoking prevalence: An empirical investigation in ten developed countries. *British Actuarial Journal*, 18(2):452–466, 2013.

Sebastian Kraemer. The fragile male. *BMJ*, 321(7276):1609–1612, 2000.

Hong Li and Yang Lu. Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin: The Journal of the IAA*, 47(2):563–600, 2017.

Jackie Li. A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population studies*, 67(1):111–126, 2013.

Nan Li and Ronald Lee. Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594, 2005.

Mike Ludkovski, Jimmy Risk, and Howard Zail. Gaussian process models for mortality rates and improvement factors. *ASTIN Bulletin: The Journal of the IAA*, 48(3):1307–1347, 2018.

Spyros Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, December 1993.

Peter Z. G Qian, Huaiqing Wu, and C. F. Jeff Wu. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics*, 50(3):383–396, 2008.

Adrian E Raftery, Nan Li, Hana Ševčíková, Patrick Gerland, and Gerhard K Heilig. Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109:13915–13921, 2012.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Jennifer C. Regan and Linda Partridge. Gender and longevity: Why do men die earlier than women? comparative and experimental evidence. *Best Practice & Research Clinical Endocrinology & Metabolism*, 27(4):467 – 479, 2013.

Olivier Roustant, David Ginsbourger, and Yves Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.

Olivier Roustant, Esperan Padonou, Yves Deville, Aloïs Clément, Guillaume Perrin, Jean Giorla, and Henry P. Wynn. Group kernels for Gaussian process metamodels with categorical inputs. working paper or preprint, July 2018. URL `https://hal.archives-ouvertes.fr/hal-01702607`.

Han Lin Shang. Mortality and life expectancy forecasting for a group of populations in developed countries: a multilevel functional data method. *The Annals of Applied Statistics*, 10(3):1639–1672, 2016.

United Nations. *World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables*. Department of Economic and Social Affairs, Population Division (2011), 2011. ST/ESA/SER.A/313.

Chou-Wen Wang, Sharon S Yang, and Hong-Chih Huang. Modeling multi-country mortality dependence and its application in pricing survivor index swaps—–a dynamic copula approach. *Insurance: Mathematics and Economics*, 63:30–39, 2015.

Arkadiusz Wiśniowski, Peter WF Smith, Jakub Bijak, James Raymer, and Jonathan J Forster. Bayesian population forecasting: Extending the Lee-Carter method. *Demography*, 52(3):1035–1059, 2015.

Sharon S Yang and Chou-Wen Wang. Pricing and securitization of multi-country longevity risk with mortality dependence. *Insurance: Mathematics and Economics*, 52(2):157–169, 2013.

About The Society of Actuaries

With roots dating back to 1889, the Society of Actuaries (SOA) is the world's largest actuarial professional organization with more than 31,000 members. Through research and education, the SOA's mission is to advance actuarial knowledge and to enhance the ability of actuaries to provide expert advice and relevant solutions for financial, business and societal challenges. The SOA's vision is for actuaries to be the leading professionals in the measurement and management of risk.

The SOA supports actuaries and advances knowledge through research and education. As part of its work, the SOA seeks to inform public policy development and public understanding through research. The SOA aspires to be a trusted source of objective, data-driven research and analysis with an actuarial perspective for its members, industry, policymakers and the public. This distinct perspective comes from the SOA as an association of actuaries, who have a rigorous formal education and direct experience as practitioners as they perform applied research. The SOA also welcomes the opportunity to partner with other organizations in our work where appropriate.

The SOA has a history of working with public policymakers and regulators in developing historical experience studies and projection techniques as well as individual reports on health care, retirement and other topics. The SOA's research is intended to aid the work of policymakers and regulators and follow certain core principles:

Objectivity: The SOA's research informs and provides analysis that can be relied upon by other individuals or organizations involved in public policy discussions. The SOA does not take advocacy positions or lobby specific policy proposals.

Quality: The SOA aspires to the highest ethical and quality standards in all of its research and analysis. Our research process is overseen by experienced actuaries and nonactuaries from a range of industry sectors and organizations. A rigorous peer-review process ensures the quality and integrity of our work.

Relevance: The SOA provides timely research on public policy issues. Our research advances actuarial knowledge while providing critical insights on key policy issues, and thereby provides value to stakeholders and decision makers.

Quantification: The SOA leverages the diverse skill sets of actuaries to provide research and findings that are driven by the best available data and methods. Actuaries use detailed modeling to analyze financial risk and provide distinct insight and quantification. Further, actuarial standards require transparency and the disclosure of the assumptions and analytic approach underlying the work.

Society of Actuaries
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org