Annals of Actuarial Science (2021), **15**, pp. 318–345
doi:10.1017/S1748499521000142

PAPER

Multi-output Gaussian processes for multi-population longevity modelling

Nhan Huynh and Mike Ludkovski*

Department of Statistics and Applied Probability, University of California at Santa Barbara, Santa Barbara, CA 93106, USA
*Corresponding author. E-mail: ludkovski@pstat.ucsb.edu

(Received 01 March 2020; revised 17 March 2021; accepted 24 March 2021; first published online 17 May 2021)

Abstract

We investigate joint modelling of longevity trends using the spatial statistical framework of Gaussian process (GP) regression. Our analysis is motivated by the Human Mortality Database (HMD) that provides unified raw mortality tables for nearly 40 countries. Yet few stochastic models exist for handling more than two populations at a time. To bridge this gap, we leverage a spatial covariance framework from machine learning that treats populations as distinct levels of a factor covariate, explicitly capturing the cross-population dependence. The proposed multi-output GP models straightforwardly scale up to a dozen populations and moreover intrinsically generate coherent joint longevity scenarios. In our numerous case studies, we investigate predictive gains from aggregating mortality experience across nations and genders, including by borrowing the most recently available "foreign" data. We show that in our approach, information fusion leads to more precise (and statistically more credible) forecasts. We implement our models in `R`, as well as a Bayesian version in `Stan` that provides further uncertainty quantification regarding the estimated mortality covariance structure. All examples utilise public HMD datasets.

Keywords: longevity modeling; Gaussian process; multi-population mortality

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 summarised in dozens of national and sub-national registries, presenting a major data-analysis challenge. The burgeoning Human Mortality Database (HMD, 2018) offers a centralised portal to nearly 40 such datasets, yielding a rich source of cross-national longevity trends.

Significant predictive 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 vis-a-vis "noise") and simultaneously reduce model risk by increasing credibility of the forecasts. Moreover, joint models capture 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.

Yet few models exist for predictive multi-population longevity analysis beyond two populations. The latter case affords the convenient hierarchy of treating one population as the baseline "index" and then modelling the "spread" between the index and the secondary population. With three or more populations, the conceptual meaning of the multiple resulting longevity spreads

© The Author(s), 2021. Published by Cambridge University Press on behalf of Institute and Faculty of Actuaries

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

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 dozens of parameters, one quickly ends up with hundreds of parameters to be estimated, creating significant computation and statistical inference bottlenecks.

In this article, we investigate a machine learning approach that simultaneously models multiple longevity surfaces within a joint *spatial covariance* framework. Our approach scales up to ~ 10 populations, which we treat as a factor covariate, with the respective correlation inferred as part of the model fitting procedure. Specifically, we employ multi-output Gaussian process (MOGP) models, building upon our earlier work in Ludkovski *et al.* (2018) and Huynh *et al.* (2020) on the use of GPs for longevity predictive analytics. GPs treat age-specific mortality rates as a noisily observed response surface that is learned via the multivariate *kriging* equations. Embedding multiple populations within a MOGP naturally captures the borrowing of mortality information and the underlying commonality of mortality patterns. Indeed, a MOGP imposes a transparent correlation structure on the co-dependence of mortality rates across populations, disentangling it from the global Age–Year pattern. Moreover, GPs afford a Bayesian perspective, yielding a full uncertainty quantification – including coherent multi-population stochastic scenarios – not just for mortality rates but also for mortality improvement factors.

The GP paradigm brings a flexible non-parametric spatio-temporal structure that treats tasks of data smoothing (aka in-sample prediction) and forecasting (aka out-of-sample prediction) in an internally consistent manner. The vast GP ecosystem has become a centrepiece of probabilistic data science and includes a multitude of extensions, from GP GLMs (to address non-Gaussian observation noise) to Kronecker GP (for faster analysis of gridded data). We refer to Duvenaud (2014), Chu & Ghahramani (2005), Garrido-Merchán & Hernández-Lobato (2018) for further discussion of GP modelling with categorical covariates. From the machine learning perspective, the respective ideas of multi-task learning and transfer learning have wide applications (Bonilla *et al.*, 2008; Caruana, 1997; Letham & Bakshy, 2019).

In the context of mortality modelling, single-population GPs were investigated in Ludkovski *et al.* (2018) and some preliminary multi-population analysis appeared recently in Huynh *et al.* (2020). Related spatio-temporal methods were investigated 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). Another way to introduce dependence between populations is through statistical shrinkage within a Bayesian hierarchical model. Raftery *et al.* (2012) modelled 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 & Cairns, 2013; Boonen & Li, 2017).

The strand of literature addressing multi-population extensions of the now-classical Lee–Carter stochastic mortality framework is getting longer. The seminal work by Li & 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) common age effect (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 & Lee (2005) functionally corresponds to having a single degree

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

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). To achieve coherent forecasts with two populations, Hyndman *et al*. (2013) considered co-integration, see also the multi-level functional regression approach in Shang (2016). In a similar spirit, D'Amato *et al*. (2016), Li & Lu (2017), and Guibert *et al*. (2019) investigated vector autoregressive approaches to achieve correlation across the multiple Period factors. Alternatively, Chen *et al*. (2015) and Wang *et al*. (2015) applied a copula approach to capture the dependence between individual Period factors and Yang & Wang (2013) considered a vector error correction model. As further extensions, Wang *et al*. (2020) incorporate structural breaks in order to capture time-varying mortality improvement factors and Dong *et al*. (2020) apply tensor decomposition to identify the univariate Age, Year and Population factors.

Let us also mention other methods for multi-population mortality, such as those based on hierarchical Buhlmann credibility theory (Tsai & Zhang, 2019; Tsai & Wu, 2020), hierarchical Bayes (Lu *et al*., 2019) and continuous-time Levy processes (Qin & Jevtic, 2016).

To sum up, our main contribution is a statistical methodology to build multi-population longevity models through MOGP. A key innovation is the use of intrinsic coregionalisation model (ICM) kernels to achieve dimension reduction that leads to more efficient models that can gracefully handle two to ten populations. Another innovation is a detailed discussion about how to pool populations to maximise predictive power. The analysis below supersedes the earlier proceedings version in Huynh *et al*. (2020) that concentrated on descriptive investigation of single-population GPs across HMD datasets and primarily focused on two-population cases with full-rank kernels. As further supplementary material, we also provide an online tutorial with additional animations and visualisations, see https://nhanhuynh46.github.io/MOGPTutorials/.

The rest of the paper is organised as follows. Section 2 describes the MOGP model for multiple-population mortality analysis. Section 3 focuses on how MOGPs can maximise predictive accuracy, while section 4 shows how MOGP is appropriate for coherent forecasting and capturing commonality of mortality experience.

1.1 Motivating multi-population models

Our first motivation for developing multi-population models is to improve predictive accuracy. It is generally accepted that there is strong commonality in mortality experiences of different populations, and therefore that there is an opportunity for data fusion to better capture trends and de-noise raw data. Hence, we wish to build a multi-population model to maximise its "credibility", or equivalently reduce the mis-specification between the true mortality evolution and the fitted model dynamics that arises due to using limited historical data. The latter idea has several complementary aspects that we now enumerate to foreshadow the methods and results below.

First, data fusion is intrinsically linked to data selection: one should only model populations that are actually dependent; including those that are little-correlated is likely to worsen model fit and forecasting performance. Therefore, multi-population modelling is closely linked to identifying dependence patterns and selecting (i.e. clustering) populations that are most correlated. From a modelling perspective, judicious choice of which populations to aggregate is critical to keeping models tractable and computationally lean – it is beyond the reach of nearly all methods to directly handle the full HMD with 75+ datasets.

Second, data fusion is also important for mitigating model risk, i.e. for fitting the best model. Therefore, successful data fusion is expected to manifest itself in reduced parameter uncertainty. For GPs, this translates into tighter hyperparameter and latent-surface posteriors, affording a transparent visualisation of higher model stability and higher confidence into the predictive forecasts. We emphasise the latter Bayesian concepts and assess our predictive accuracy not simply

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

through the predictive mean (via a mean-squared-error related metric) but also through scoring rules, such as continuous ranked probability score (CRPS) that are based on the quality of the full predictive distribution.

Third, for HMD data in particular, a key problem is to obtain highest-quality contemporary forecasts, i.e. to assess the present-date mortality in a given "domestic" population. Because the database is updated very frequently and asynchronously across different countries, at any given timepoint the database is not rectangular but "notched", i.e. it ends at different years. For instance, as of February 2020 some countries already have 2018 data added, most have data up to 2017, and a few are still lagging and only have data up to 2016. Using such input data to build the best-feasible prediction of 2019 mortality in say UK is possibly the most common use of HMD, but presents challenges relative to the "classical" time-series models. As we show, GPs are both easily adaptable and internally consistent for this task.

To conclude this section, we reiterate the big picture perspective that underlies the premise of multi-population analysis: because populations *are* similar – both in their static structure and in their dynamic evolution – one should leverage this similarity to improve predictive analysis. But this must be done in a smart way, keeping in mind computational and statistical considerations and the specific predictive tasks envisioned. We believe that for mortality analysis, a multi-output machine learning framework checks off all these boxes and offers a tractable and scalable way to jointly analyse a collection of two to ten populations.

Data source: We work with mortality data from the 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 and presently focuses on developed economies where death registrations and census data are available and reliable. For our analysis, we rely on 1-year age groups, concentrating on Ages 50–84 (retirement ages most relevant for predictive actuarial analysis) for both genders and calendar Years 1990–2016. In the models below, we consider various subsets of the following 16 European datasets: Austria (AUT), Belgium (BEL), Czech Republic (CZE), Denmark (DEN), Estonia (EST), France (FRA), Germany (GER), Hungary (HUN), Latvia (LAT), Lithuania (LTU), Netherlands (NED), Poland (POL), Spain (ESP), Sweden (SWE), Switzerland (SUI) and United Kingdom (UK or GBR).

The dataset is organised as a large table. The nth observation for the lth country 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.

2. MOGP Models

2.1 GP regression for mortality tables

A 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 characterised 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:

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

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

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

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 reflect 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 materialise in the absence of any random shocks. We assume that observation noise is Gaussian: $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$ or $\epsilon = (\epsilon^1, ..., \epsilon^n) \sim \mathcal{N}(\mathbf{0}, \Sigma = \text{diag}(\sigma^2))$. It follows that $\text{Cov}(y^i, y^j) = \text{Cov}(f(x^i), f(x^j)) + \sigma^2 \delta(x^i, x^j)$ and therefore $\mathbf{y} \sim \mathcal{N}(\mathbf{m}(\mathbf{x}), \mathbf{C}(\mathbf{x}, \mathbf{x}) + \Sigma)$ where $\delta(x^i, x^j)$ is the Kronecker delta.

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 $p(\mathbf{f}|\mathcal{D}) \propto p(\mathbf{f})p(\mathbf{y}|\mathbf{x}, \mathbf{f})$:

$$\text{Posterior distribution } = \frac{\text{Prior distribution x Likelihood function}}{\text{Marginal distribution}}$$

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:

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

where $\mathbf{C}(\mathbf{x}, \mathbf{x}_*)$ is the covariance matrix between $f(\mathbf{x})$ and $f(\mathbf{x}_*)$, $\mathbf{C}_{**}$ is the covariance matrix of $f(\mathbf{x}_*)$, $\Sigma_{**} = \text{diag}(\sigma^2)$ is the noise variance matrix of the test inputs $\mathbf{x}_*$ and $\mathbf{m}_* = m(\mathbf{x}_*)$. The MVN formulas then imply that

$$p(\mathbf{y}_*|\mathbf{y}) \sim \mathcal{N}(\mathbf{m}_*(\mathbf{x}_*), \mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*)) \qquad \text{where}$$

$$\mathbb{E}[\mathbf{y}_*|\mathcal{D}] = \mathbf{m}_*(\mathbf{x}_*) = \mathbf{m} + \mathbf{C}(\mathbf{x}, \mathbf{x}_*)^T [\mathbf{C} + \Sigma]^{-1} (\mathbf{y} - \mathbf{m}); \tag{3}$$

$$\mathbb{V}ar(\mathbf{y}_*|\mathcal{D}) = \mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*) = \mathbf{C}_{**} + \Sigma_{**} - \mathbf{C}(\mathbf{x}, \mathbf{x}_*)^T [\mathbf{C} + \Sigma]^{-1} \mathbf{C}(\mathbf{x}_*, \mathbf{x}) \tag{4}$$

Note that the posterior variance $\mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*)$ is equal to the prior variance $\mathbf{C}_{**} + \Sigma_{**}$ minus a positive term which reflects the information gained (relative to the prior) from the training data. Furthermore, (3)–(4) are valid for *any* $\mathbf{x}_*$, i.e. both for in-sample smoothing or for out-of-sample extrapolation.

Covariance function. The kernel $C(x, x')$ 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 (SE) kernel:

$$\tilde{C}(x^i, x^j) = \eta^2 \exp\left[-\frac{(x^i_{ag} - x^j_{ag})^2}{2\theta^2_{ag}} - \frac{(x^i_{yr} - x^j_{yr})^2}{2\theta^2_{yr}} \right] \tag{5}$$

When $x^i \approx x^j$, the covariance reaches its maximum value η^2; when x^i and x^j are far apart, $\tilde{C}(x^i, x^j) \approx 0$. 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 (5) that are called characteristic length-scales. The length-scales determine how much influence an observation has on others in the Age and Year dimensions, respectively. Note that θ_{ag} – length-scale for Age – and θ_{yr} – length-scale for Year – are not comparable. An important aspect that influences the goodness-of-fit of a GP model is its spatial smoothness. The SE covariance kernel (5) 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

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

be exploited below for computing mortality improvement factors. Moreover, the length-scales θ affect the qualitative nature of the fitted $m_*(\cdot)$. When length-scales are too large, the fitted curves are over-smoothed and the influence of individual data points attenuates (Rasmussen & 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 length-scales indicate over-fitting of the spatial dependence, generating high-frequency oscillations in the fitted $m_*(\cdot)$ and low observation noise σ^2.

2.2 MOGP kernels

The idea of commonality in mortality experiences is equivalent to the existence of global longevity trends. In the context of a spatio-temporal model, it implies an underlying *shared* covariance structure. This can be easily verified visually or statistically, see Huynh *et al.* (2020) for comparison of 10+ European countries available in HMD where we observe both similar fitted mortality dynamics (e.g. similar mortality improvement curves over time) and similar estimated GP covariance hyperparameters.

Let L be the number of different populations considered. To jointly model the L different outputs, $\{f_l\}_{1\le l\le L}$ we correlate them using the framework of MOGPs which was introduced in geostatistics under the name of multivariate kriging or co-kriging (Chiles & Delfiner, 1999; Hoef & Barry, 1998). The aim of co-kriging is to estimate the under-sampled variables using spatial correlation with other sampled variables.

The vector-valued latent response over the Age–Year input space is defined as:

$$\mathbf{f}(\mathbf{x}) = (f_1(x^1), \ldots, f_1(x^N), \ldots, f_L(x^1), \ldots, f_L(x^N)) = (f_1(\mathbf{x}), \ldots, f_L(\mathbf{x}))$$

where the functions $\{f_l\}_{l=1}^L$ are the log-mortality surface for the corresponding population l. Similar to single-output GP (SOGP), MOGP assumes that the vector-valued $\mathbf{f}$ follows a GP:

$$\mathbf{f} \sim GP(\mathbf{m}, \mathbf{C})$$

where $\mathbf{m} \in \mathbb{R}^{LN\times 1}$ is the mean vector whose elements $\{m_l(\mathbf{x})\}_{l=1}^L$ are the mean functions of each output, and $\mathbf{C} \in \mathbb{R}^{LN\times LN}$ is the fused covariance matrix.

Populations are treated as categorical input, encoded via L additional input dimensions with 0/1 encoding. Thus, the 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 $x^n_{pop,l} = \mathbb{1}_{\{\text{population}=l\}}$ are indicators, set to 1 if and only if the nth observation is from population l. We denote by N_l the number of Age–Year inputs for population l. If all L populations have the same set of inputs and $N_1 = \ldots = N_L = N$, the dataset is said to be *isotropic*.

To construct the fused $\mathbf{C}$, one approach is to take the product between a kernel for the Age–Year covariates x_{ag}, x_{yr} and a kernel for the categorical ones (Qian *et al.*, 2008; Roustant *et al.*, 2018). Let:

$$\Gamma_{(l_1,i),(l_2,j)} = \exp\left[-\theta_{l_1,l_2}\delta^{ij}_{l_1,l_2}\right] \quad \text{where} \quad l_1, l_2 \in \{1, \ldots, L\} \tag{6}$$

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.

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

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

$$C(x^i, x^j) := \eta^2 \exp\left[-\frac{(x^i_{ag} - x^j_{ag})^2}{2\theta^2_{ag}} - \frac{(x^i_{yr} - x^j_{yr})^2}{2\theta^2_{yr}} \right] \prod_{\{l_1,l_2\}} \exp\left[-\theta_{l_1,l_2} \delta^{ij}_{l_1,l_2} \right] \tag{7}$$

$$= \begin{cases} \tilde{C}_{i,j} & \text{if observations are from the same population;} \\ \tilde{C}_{i,j}\Gamma_{(l_1,i),(l_2,j)} & \text{if observations are from populations } l_1, l_2 \end{cases}$$

When observations are from the same country, the covariance between the ith and jth observation is the same as in a SOGP model, cf. equation (5). Intuitively, $\Gamma_{l_1,l_2} < 1$ discounts the covariance when observations are from different populations and is driven by the hyperparameter θ_{l_1,l_2}: large value of θ_{l_1,l_2} implies low correlation $r_{l_1,l_2} := \exp(-\theta_{l_1,l_2})$ between the two populations.

Two important assumptions are made in equation (7). First, there is separability (Alvarez *et al.*, 2011) between the cross-population covariance and the covariance over the Age–Year inputs. Second, observations across L populations share the same spatial covariance kernel. This assumption is useful to examine the commonality in the mortality across populations via the length-scales in Age and Year dimensions. It can be thought of as full statistical shrinkage of the individual population covariances towards a common baseline, cf. section 4.1 below.

We note that unlike the Li & Lee (2005) framework and its extensions that explicitly assume a common factor shared by all populations, the GP framework is more similar to co-integration frameworks that postulate statistical dependence across f_ℓ's. In that sense, GP modelling is more amenable to handling larger collections of populations where more complex mortality clustering may be present, see section 3.1.

2.3 Coregionalised kernels

Estimating cross-population covariance in (7) requires fitting $L(L-1)/2$ parameters $\theta_{l_1,l_2}, 1 \le l_1 \le l_2 \le L$ which imposes challenges in both statistical and computational aspects when modelling MOGP with many outputs (e.g. $L \ge 4$). An attractive dimension reduction approach to keep the number of correlation parameters low is the ICM (Alvarez *et al.*, 2011). In ICM, each output f_l is assumed be a linear combination of independent latent GPs. Let $u_1(\mathbf{x}), \ldots, u_Q(\mathbf{x})$ be independent latent functions each from a GP prior with covariance kernel $C^{(u)}(\mathbf{x}, \mathbf{x}')$. The modelled outputs f_l are linear combinations of these Q latent factors:

$$f_l(\mathbf{x}) = a_{l,1}u_1(\mathbf{x}) + \ldots + a_{l,Q}u_Q(\mathbf{x}) = \sum_{q=1}^{Q} a_{l,q}u_q(\mathbf{x}) \tag{8}$$

where $a_{l,q}$'s are the factor loadings. Let $\mathbf{a}_q = (a_{1,q}, \ldots, a_{L,q})^T$ be the vector representing the collection of linear coefficients associated with the qth latent function across the L outputs, so that $\mathbf{f(x)} = \sum_{q=1}^{Q} \mathbf{a}_q u_q(\mathbf{x})$. It follows that the covariance for $\mathbf{f(x)}$ is

$$\begin{aligned} \mathbf{C}(x, x') = \mathrm{Cov}\left(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')\right) &= \mathrm{Cov}\left(\sum_{q=1}^{Q} \mathbf{a}_q u_q(\mathbf{x}), \sum_{q=1}^{Q} \mathbf{a}_q u_q(\mathbf{x}') \right) \\ &= \left(\sum_{q=1}^{Q} \mathbf{a}_q \mathbf{a}_q^T \right) \otimes \mathrm{Cov}\left(u_q(\mathbf{x}), u_q(\mathbf{x}')\right) \\ &= AA^T \otimes C^{(u)}(\mathbf{x}, \mathbf{x}') \end{aligned} \tag{9}$$

where matrix $A = (\mathbf{a}_1, \ldots, \mathbf{a}_Q)$ and $\otimes$ is the Kronecker matrix product. Re-parametrising by $B := AA^T$, (9) can be expressed as the Kronecker product between the cross-population covariance

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

$B \in \mathbb{R}^{L\times L}$ and the covariance over the Age–Year inputs $\mathbf{C}^{(u)} \in \mathbb{R}^{N\times N}$:

$$\mathbf{C} = \mathrm{Cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) = B \otimes \mathbf{C}^{(u)} \tag{10}$$

The *coregionalisation matrix* B has rank Q. Under ICM, the number of hyperparameters in the cross-population covariance is reduced from $L(L-1)/2$ to $Q \times L$. Hence, taking $Q < L/2$ allows to reduce the hyperparameter space and alleviate the computational budget compared to the "full rank" setup.

ICM is a special form of the linear coregionalisation model (LCM) (Alvarez *et al.*, 2011). In LCM, the independent latent functions are from different GP priors to capture all possible variability from multiple outputs. Its application is beyond the scope of this paper. Notably, the assumption in ICM that all L populations share the common spatial covariance suits our interest in the inference of joint length-scales in Age and Year dimensions. The computational complexity required in ICM is greatly simplified due to the properties of the Kronecker product. Finally, ICM allows the process variance η_l^2 to vary by populations, i.e., the l*th* element in the diagonal in B is the process variance for population l. This further bolsters the flexibility in MOGP to excel in out-of-sample forecasts.

The Kronecker decomposition in (10) is also highly useful to speed up the overall fitting. The most expensive step in building a MOGP is solving for the inverse of the covariance matrix $\mathbf{C}$ that shows up in the log-likelihood function. Indeed, inverting this $LN \times LN$ matrix is computationally expensive even for modest values of L and N. Since the inverse of the Kronecker product is equal to the product of the respective inverses: $\left(B \otimes C^{(u)}\right)^{-1} = B^{-1} \otimes \left(C^{(u)}\right)^{-1}$, the ICM structure reduces the computational complexity from $\mathcal{O}((LN)^3)$ to $\mathcal{O}(L^3 + N^3 + LN)$. In our experience, this translates to a factor of 2–4 speed-up in computational time, allowing us to scale up to 10–12 populations. We emphasise that the overall MOGP (and ICM) structure is completely agnostic to L, so that exactly the same numerical method is applied to handle $L = 2$ populations or $L = 10$.

Selecting ICM rank *Q*. Because Q is not one of the hyperparameters to be optimised, we exploit the Bayesian information criterion (BIC) to select the value of Q for the most parsimonious models. To illustrate the role of the rank Q of an ICM kernel, Table 1 compares several MOGP models with different $Q = 2, 3, 4$. We consider two case studies, both with $L = 8$ but with different constituent populations. First, we note the remarkably faster training time (3-time speedup) in ICM relative to a full-rank kernel; notably the speedup is independent of Q and is driven by the Kronecker matrix algebra. Second, we note that most ICM models have better predictive performance (see sections 2.7 and 3.1 below for explanation of symmetric mean absolute percentage error (SMAPE) and CRPS) than the kernel from (7). Third, BIC criterion suggests that $Q = 2$ is the preferred model in both cases. Note that there are only $2L = 16$ hyperparameters in the resulting ICM kernel, almost twice as few as $L(L-1)/2 = 28$ hyperparameters in the respective (7).

Non-rectangular datasets. We have discussed the use of ICM for isotropic data. The ICM framework can be extended to deal with partially heterotropic data where only a portion of L inputs are available and which arises in HMD due to missing data especially at the most recent years. Let M' be the number of distinct inputs across L populations and $M = N_1 + \ldots + N_L$ be the number of observations in training data. We consider the setting that $M' < ML$ so that for some inputs not all L outputs are observed. Define the vector-valued "complete data" function $\mathbf{f}(\mathbf{x})$, with $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{LM'\times 1}$. We further introduce $\mathbf{f}^o(\mathbf{x})$ as the vector-valued function corresponding to the observed outputs, $\mathbf{f}^o(\mathbf{x}) \in \mathbb{R}^{M\times 1}$. The relation between $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}^o(\mathbf{x})$ is formulated through the "communication" matrix S, $\mathbf{f}^o(\mathbf{x}) = S^T\mathbf{f}(\mathbf{x})$, where $S \in \mathbb{R}^{LM'\times M}$. The column vectors in S are orthonormal with values of 0 and 1 to eliminate the unobserved outputs, see Skolidis & Sanguinetti (2011). Applying linear transformation to a MVN vector, we can then identify the distribution of $\mathbf{f}^o(\mathbf{x})$ as a GP with covariance:

$$\mathrm{Cov}(\mathbf{f}^o(\mathbf{x}), \mathbf{f}^o(\mathbf{x}')) = S^T\mathrm{Cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}'))S = S^T(B \otimes K)S$$

recovering the Kronecker structure

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Table 1. Comparison between full-rank and ICM MOGP models. Improvement in SMAPE and CRPS uses Hungary as the target population and compares MOGP with respect to SOGP Hungary model. The reported percentages are averages over 1-year ahead Hungarian mortality forecasts for Ages 70–84 based on three training sets: 1990–2013 (predict 2014), 1990–2014 (predict 2015) and 1990–2015 (predict 2016) for same Ages 70–84. Best metrics are in *italics*

	Full rank	ICM ($Q = 2$)	ICM ($Q = 3$)	ICM ($Q = 4$)
# Kernel hyperparameters	28	16	24	32
Case study I: AUT, DEN, EST, GER, HUN, LTU, CZE, POL				
Running time (mins)	132.75	59.51	59.41	58.86
Improvement in SMAPE (%)	−2.89	*10.75*	−1.54	1.93
Improvement in CRPS (%)	2.79	*16.82*	11.09	16.63
Total BICs	–	*−28, 073*	−27, 719	−28, 006
Case study II: EST, HUN, LTU, NED, POL, SWE, SUI, GBR				
Running time (mins)	176.78	59.03	58.99	58.98
Improvement in SMAPE (%)	−24.90	−2.77	7.71	−32.76
Improvement in CRPS (%)	−2.28	13.53	14.04	−2.95
Total BICs	–	−28, 890	−28, 346	−28, 361

2.4 GP hyperparameters

To implement a GP model requires specifying its hyperparameters. Note that actual inference reduces to linear-algebraic formulas in (12)–(13), and the modelling task is to learn the spatial covariance, namely the mean and kernel functions.

2.4.1 Mean function

To capture the long-term longevity features, such as higher mortality at higher ages, we fit a parametric prior mean:

$$m(x) = \beta_0 + \sum_{j=1}^{p} \beta_j h_j(x)$$

where h_j's are given basis functions and the β_j's are unknown coefficients. The coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are estimated simultaneously with other hyperparameters. Let $\mathbf{h}(x) = \big(h_1(x), \ldots, h_p(x)\big)$, $\mathbf{H} = \big(\mathbf{h}(x^1), \ldots, \mathbf{h}(x^N)\big)$ and $\mathbf{D} = (\mathbf{C} + \Sigma)^{-1}\mathbf{H}$, then the posterior mean of $\boldsymbol{\beta}$ along with the predicted posterior mean $m_*(x_*)$ and respective variance $s_*^2(x_*)$ for a new input x_* are

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T\mathbf{D})^{-1}\mathbf{H}^T(\mathbf{C} + \Sigma)^{-1}\mathbf{y}; \tag{11}$$

$$m_*(x_*) = \mathbf{h}(x_*)^T\hat{\boldsymbol{\beta}} + \mathbf{c}(x_*)^T(\mathbf{C} + \Sigma)^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}); \tag{12}$$

$$s_*^2(x_*) = C(x_*, x_*) + (\mathbf{h}(x_*)^T - \mathbf{c}(x_*)^T\mathbf{D})^T(\mathbf{H}^T\mathbf{D})^{-1}(\mathbf{h}(x_*)^T - \mathbf{c}(x_*)^T\mathbf{D}) \tag{13}$$

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 modelling 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.

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Within a multi-population model, we use a linear mean function to take into account the different trends across populations:

$$m(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \sum_{l=2}^{L} \beta_{pop,l} x_{pop,l}^n \tag{14}$$

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 (14) implies the *same* shared Age structure – mortality rates rising exponentially in x_{ag} with slope β_1^{ag} in all populations.

2.4.2 Observation likelihood

We assume a constant observation noise within each population $\sigma_l = \text{StDev}(\epsilon^i)$ for all i in (2) where $x_{pop,l}^i = 1$. This accounts for heterogeneous characteristics when observations from multiple populations are combined, in particular σ_l is smaller for larger populations (Huynh *et al.*, 2020). The σ_l's are estimated via maximum likelihood or Markov Chain Monte Carlo along with all other hyperparameters. While assuming homogeneity of noise variance in terms of Age and Year is not entirely realistic, based on the discussion in Ludkovski *et al.* (2018), the impact of more complex observation models is minimal. A common alternative is to assume a Poisson likelihood; however it is well known that mortality data are overdispersed, so a Poisson parametrisation is also mis-specified.

2.4.3 Estimating the parameters

In single-population GP, the set of hyperparameters is $\Theta = (\theta_{ag}, \theta_{yr}, \eta^2, \sigma^2, \boldsymbol{\beta})$. We can learn values of the hyperparameters via optimisation 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 $p(\mathbf{y}|\mathbf{x}, \Theta) = \mathcal{N}(\mathbf{m}, \mathbf{C} + \Sigma)$ and if we assume the mean function is known or fixed, the log-likelihood of the marginal is simply a MVN density:

$$\log p(\mathbf{y}|\mathbf{x}, \Theta) = -\frac{1}{2}\mathbf{y}^T(\mathbf{C} + \Sigma)^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{C} + \Sigma| - \frac{N}{2}\log(2\pi) \tag{15}$$

Thus, we have to solve a system of nonlinear equations to maximise (15) which yields the maximum likelihood estimate (MLE). We implement SOGP 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 optimisation algorithm. In MOGP with ICM kernel, the hyperparameters are $\Theta = (\theta_{ag}, \theta_{yr}, (a_{l,q}), (\sigma_l^2), \boldsymbol{\beta})$. We use the `R` package `kergp` (Deville *et al.*, 2019) to carry out the respective MLE via Kronecker decompositions.

2.5 Bayesian GP regression

The GP hyperparameters summarise 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

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

improper uniform priors. In our analysis, we employ weakly informative priors to minimise influence of a priori assumptions (so that the data speaks for itself) but still regularise 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 optimisation for statistical models.

Following `Stan` recommendations, we standardise 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 standardised 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 length-scales in the covariance kernel: $\theta_{ag} = \sigma_{\mathbf{x}_{ag}}\theta^{std}_{ag}$ and $\theta_{yr} = \sigma_{\mathbf{x}_{yr}}\theta^{std}_{yr}$.

2.5.1 Bayesian versus MLE MOGP

To illustrate uncertainty quantification of a MOGP using a Bayesian `Stan` framework, we build a joint model on four Male populations from Denmark, France, Sweden and UK. The MOGP model uses the full-rank kernel (7) trained on Ages 70–84 and Years 1990–2012. For Bayesian 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: $\theta^{std}_{ag} \sim$ Inv-Gamma(9, 12), $\theta^{std}_{yr} \sim$ 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 $\log \eta^2 \sim \mathcal{N}(-3, 1)$, and for observation noise $\sigma^2 \sim \mathcal{N}_+(0, 0.5)$. Finally, for the population length-scales in (7), we use $\log \theta_{l_1,l_2} \sim \mathcal{N}(-1, 1)$ for all l_1, l_2. (Implementing a `Stan` model for ICM kernels is beyond the scope of this work.)

Table B.1 in the Appendix reports all the resulting hyperparameters using the `kergp` engine in `R` and the `Stan` HMC. We observe that all MLE fits are within the 95% posterior credible intervals from the `Stan` model. Also, the 95% credible interval for β_1^{ag} confirms the significance of the linear effect of Age. Treating Denmark as the baseline country in the mean function, the 95% CIs of all coefficients $\beta_{pop,l}$'s contain 0, implying that the differences in mortality between Denmark and other populations 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.

2.6 Kernel selection

The choice of the kernel family is the centrepiece of GP modelling. In this article, we employ the SE kernel, but manifold alternatives exist. Investigating which kernel is best is beyond the scope

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

of this paper which focuses on how to capture the multi-population aspect, while kernel selection is best analysed within a single-population setting. Indeed, our main construction in (7) relies on multiplying the single-output kernel $\tilde{C}$ with the cross-population covariance structure Γ, so that one can straightforwardly substitute *any* $\tilde{C}$.

To this end, let us briefly discuss three useful ways to do so. First, one may consider other kernel functions, such as the Matérn family (Rasmussen & Williams, 2005); in particular Matérn-5/2 is a common default choice in a machine learning context. Compared to SE which yields infinitely differentiable $m(\cdot)$, Matérn-5/2 gives predictive means that are only twice differentiable. In particular, this implies that the respective MI would be just once-differentiable and might exhibit higher-order discontinuities. Empirically, that leads to more "wiggly" fits that visually fit the historical data closer. Because mortality is expected to change slowly and smoothly, such behaviour might not be desirable, although a comprehensive analysis is needed to assess which kernel works better.

Second, one may employ *composite* kernels, i.e. go beyond the simple form of the SE kernel through adding and multiplying several covariance functions. We refer to Duvenaud (2014) for a comprehensive guide. For instance, one may embed a linear trend in Age directly inside the covariance function by taking

$$\tilde{C}(x^i, x^j) = \alpha_0^2 + \alpha_{ag}^2 x_{ag}^i x_{ag}^j + \eta^2 \exp\left[-\frac{(x_{ag}^i - x_{ag}^j)^2}{2\theta_{ag}^2} - \frac{(x_{yr}^i - x_{yr}^j)^2}{2\theta_{yr}^2} \right] \tag{16}$$

This allows to set the priors of the mean parameters within the covariance kernel.

Third, one may design *custom* covariance functions that target special features of mortality surfaces. For example, it is well known that many populations exhibit a birth cohort effect, whereby mortality is a function not just of Age and Year, but also of Birth Year = Year − Age. Thus, one could consider a kernel that incorporates dependence on Birth Year. One example is

$$\tilde{C}(x^i, x^j) = \eta^2 \exp\left[-\frac{(x_{ag}^i - x_{ag}^j)^2}{2\theta_{ag}^2} - \frac{(x_{yr}^i - x_{yr}^j)^2}{2\theta_{yr}^2} - \frac{(x_{yr}^i - x_{ag}^i - x_{yr}^j + x_{ag}^j)^2}{2\theta_{bc}^2} \right] \tag{17}$$

This would work exactly like the analysis below, except one must fit three length-scale hyperparameters θ. Note that because the covariates enter into the covariance function nonlinearly, there is no issue with parameter identification.

As mentioned, all the above generalisations would be computationally trivial to add. Some effort might be necessary to obtain the analytic expressions for mortality improvement (MI, i.e. the gradient of $\tilde{C}$). Rigorous model selection would require to focus on a given performance metric (such as SMAPE) and would also need to account for the varying model complexity. Appendix C provides illustrations of employing alternative kernel families, namely a Matern-5/2 kernel, an additive linear + SE kernel and the Birth Cohort kernel from (17). Based on numerical experiments, we find that all the above modifications play a second-order role in the quality of the model.

2.7 Performance metrics

To assess model performance we employ two metrics. First, we consider out-of-sample predictive accuracy, comparing realised 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 fact that mortality errors will be necessarily larger at higher Ages due to smaller exposed cohorts. To remedy this, we employ the mean absolute percentage error (MAPE) metric, specifically its symmetric (SMAPE) version that corrects for the tendency of MAPE to put heavier penalties on over-estimating the observations (Armstrong & Collopy, 1992; Makridakis, 1993):

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

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

where y_*^i is the realised 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 datasets.

In addition to SMAPE, we also use the CRPS metric to assess the quality of the probabilistic forecasts produced by a MOGP. Indeed, one of the major benefits of GP-based mortality models is a full distribution for future observations $y_*(x_*)$ which allows a more detailed uncertainty quantification beyond just looking at the predictive mean $m_*(x_*)$. CRPS is an example of a proper scoring rule and is defined as

$$\text{CRPS}(F, y_*) := \int_{\mathbb{R}} \left(F(z) - \mathbb{1}_{\{z \geq y_*\}}\right)^2 \, dz \tag{19}$$

where F is the predictive (cumulative) distribution and y_* is the realised outcome. Averaging over many outcomes, CRPS can be interpreted as the squared difference between the forecasted and the empirical cumulative distribution functions. In particular, CRPS penalises both bias and excessive predictive variance. A model with lower mean CRPS is judged to be better.

Mortality improvement factors. 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:

$$\partial m_{back}^{GP}(x) := \left[1 - \frac{\exp\left(m_*(x_{ag}; x_{yr})\right)}{\exp\left(m_*(x_{ag}; x_{yr}-1)\right)}\right] \tag{20}$$

3. Maximising Predictive Accuracy through Joint Modelling

We begin our illustrations by showcasing the improved predictive accuracy available from fusing data from two populations. Our first case study includes Male mortality modelling across Sweden and Denmark ($l = 1$: Denmark and $l = 2$: Sweden). The two countries share similar demographic characteristics and are Nordic neighbours. Our mean function takes into account the separation in mortality between Denmark and Sweden:

$$m(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \beta_{pop,2} x_{pop,2}^n \tag{21}$$

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. Our second case study looks at joint Male/Female modelling for Denmark; in that case $\beta_{pop,2} \equiv \beta_{FEM}$ is the mean difference between female log-mortality and male log-mortality.

Figure 1 shows the raw observations together with the GP-based predictive intervals for the first case study. Specifically, we plot the 95% predictive credible bands for $y_*(x_*)$ for three representative ages. The forecast period includes both in-sample (1990–2012) and out-of-sample (2013–2020). We observe that while for Denmark the differences between SOGP and MOGP forecasts are very slight, in Sweden the two forecasts differ noticeably out-of-sample. Table 2 compares the predictive accuracy between the models and indicates that the MOGP forecast is more accurate (smaller SMAPE) in both populations.

To highlight further differences between SOGP and full MOGP, Figure 2 examines the respective predicted annual mortality improvement factors $\partial m_{back}^{GP}(x)$. We compare SWE and DEN SOGPs against a two-population DEN+SWE MOGP, and a four-population

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Table 2. Prediction accuracy via SMAPE for SOGP and two-population full-rank models. The test set is Ages 70–84 in Years 2013, 2015 and 2016. More accurate predictions with smaller SMAPE values are in boldface.

SMAPE			2013 (1-year out)		2015 (3-year out)		2016 (4-year out)	
			SOGP	MOGP	SOGP	MOGP	SOGP	MOGP
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**
			SOGP	MOGP	SOGP	MOGP	SOGP	MOGP
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**

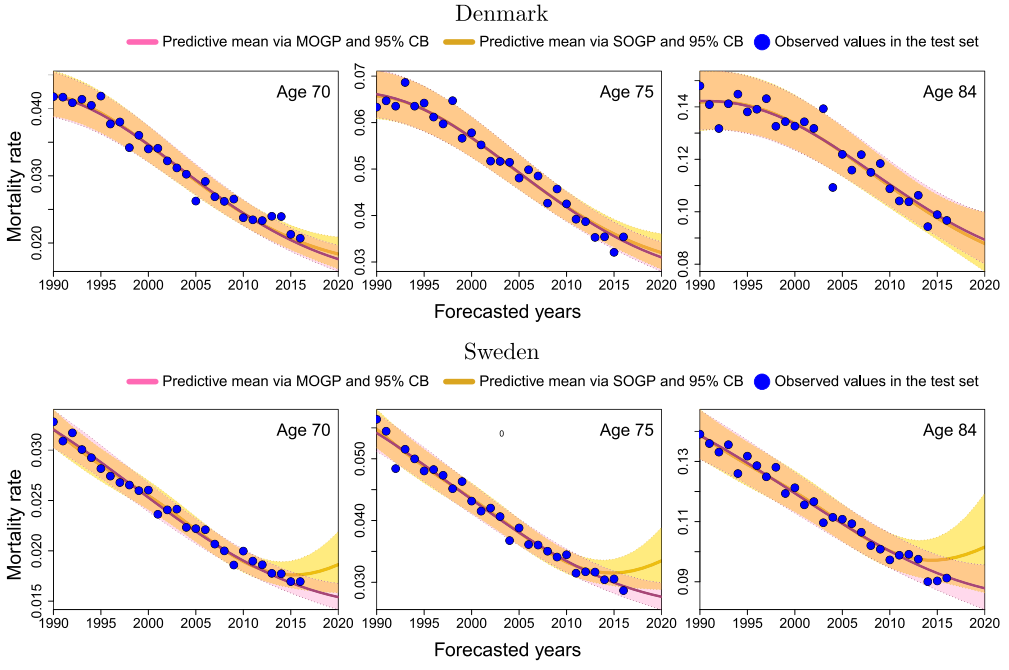


Figure 1. 95% credible intervals for observed log-mortality $y(x_*)$ across the individual SOGP and full MOGP models. Top row: Denmark Males; bottom row: Sweden Males. Up to 2011, the smoothed mortality curves and CIs are essentially identical.

DEN+FRA+GBR+SWE MOGP model, cf. Table B.1 in the Appendix. Large θ_{ag} length-scales in SOGP models lead to essentially linear improvement rate factors (blue curves). In the SWE + DEN MOGP (green curves), the Age length-scale decreases and $s_*(x_*)$ falls, so the improvement rate factors become more Age-dependent and with tighter credible bands. This effect becomes even stronger with four populations. The corresponding smoothed curves (coloured 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-population model would often smooth latest changes away.

3.1 Selecting populations for a joint model

Intuitively, incorporating more information from different populations through a MOGP ought to produce more accurate predictions and reduce predictive uncertainty. To visualise

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Table 3. Prediction quality via 3-year average improvement in SMAPE and CRPS in ICM models with $L = 2, \ldots, 6$ populations. The baseline models are SOGP for UK and Hungary, respectively

	Improvement in SMAPE (%)	Improvement in CRPS (%)		Improvement in SMAPE (%)	Improvement in CRPS (%)
UK + 1	−2.805	1.834	Hungary + 1	7.501	8.293
UK + 2	1.189	3.218	Hungary + 2	1.839	8.543
UK + 3	3.995	3.279	Hungary + 3	3.913	7.218
UK + 4	1.778	0.933	Hungary + 4	2.677	15.559
UK + 5	1.764	−4.271	Hungary + 5	10.585	13.345

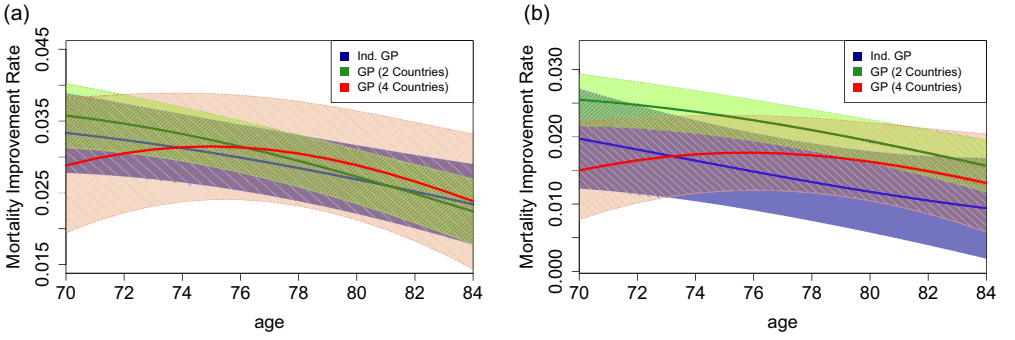


Figure 2. Comparison of annual mortality improvement factors between different joint models. Besides the mean improvement factors $\partial m^{GP}_{back}(ag;2012)$ (20) for Ages 70, . . . , 84, we also show the respective 95% posterior credible band. (a) Denmark Males. (b) Sweden Males.

how increasing L affects the changes in SMAPE and CRPS, Table 3 reports the 3-year average improvement in these metrics as L varies from 2 to 6. (ICM ranks $Q = Q(L)$ were selected each time using BIC.) Overall, we observe that information fusion is very helpful for Hungary, but not as much for UK. This links to the respective credibility of the target populations: observation noise σ_l^2 is large in Hungary but low in UK, so additional data will benefit the former more.

From a complementary perspective, Table 4 shows how the inferred cross-population correlations change as we add a new populations. We report results both for a full-rank MOGP and for ICM with $Q = 2, 3, 4$. We observe that the correlation matrix is generally stable, although some correlations can be quite different moving from one rank to another. In the AUT-SUI-GBR model, the correlation between Switzerland and UK is $r_{SUI,GBR} = 0.76$ when $Q = 2$, but rises to 0.88 when $Q = 3$. We further note a broad agreement between the correlation structure learned with an ICM and full-rank MOGP kernels.

The factor loadings ($a_{l,q}$'s) in ICM provide insight regarding the dependence patterns across populations. The interpretation of ICM loadings is analogous to principal component analysis when attempting to describe the data through independent transformed latent functions. For example, we consider factor loadings in case study II in Table 1. The best ICM kernel has rank $Q = 2$ suggesting that two latent factors are sufficient to explain variation over the eight countries considered. In fact, the first latent component is strongly correlated with Eastern-Central European countries while the second factor is the major contributor to the Western European population. This interpretation helps us identify two well-separated clusters among these eight countries, visualised in Figure 3 by plotting the ICM loadings $a_{l,2}$ against $a_{l,1}$. Note that these factor loadings can be translated into correlation and imply that Hungary is more correlated with members in the same cluster and less correlated with Western European populations.

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Table 4. Cross-corelation among two to four populations in MOGP models. All models are fitted on Ages 70–84 and reported values are averages over 3 training sets covering 1990 through 2014–2016. Italics indicate the ICM model with the smallest BIC

	Full rank	ICM ($Q=2$)	ICM ($Q=3$)	ICM ($Q=4$)
Austria, UK				
$r_{AUT,\,GBR}$	0.8432	0.8612		
Switzerland, UK				
$r_{SUI,\,GBR}$	0.8535	0.8645		
Austria, Switzerland, UK				
$r_{AUT,\,SUI}$	0.9151	0.9677	*0.9680*	
$r_{AUT,\,GBR}$	0.8590	0.8968	*0.8460*	
$r_{SUI,\,GBR}$	0.8514	0.7585	*0.8841*	
Austria, Germany, Switzerland, UK				
$r_{AUT,\,GER}$	0.9570	0.9991	*0.9869*	0.9888
$r_{AUT,\,SUI}$	0.9280	0.9956	*0.9639*	0.9658
$r_{AUT,\,GBR}$	0.8730	0.8330	*0.8917*	0.8827
$r_{GER,\,SUI}$	0.9047	0.9986	*0.9395*	0.9447
$r_{GER,\,GBR}$	0.8504	0.8091	*0.8340*	0.8267
$r_{SUI,\,GBR}$	0.8674	0.7792	*0.9047*	0.9187

	$a_{l,1}$	$a_{l,2}$
$l=1$: EST	**0.1925**	0.0815
$l=2$: HUN	**0.1692**	0.0673
$l=3$: LTU	**0.1757**	0.0411
$l=4$: NED	0.0846	**0.1127**
$l=5$: POL	**0.1670**	0.0886
$l=6$: SWE	0.0614	**0.0943**
$l=7$: SUI	0.0745	**0.1113**
$l=8$: GBR	0.1118	**0.1253**

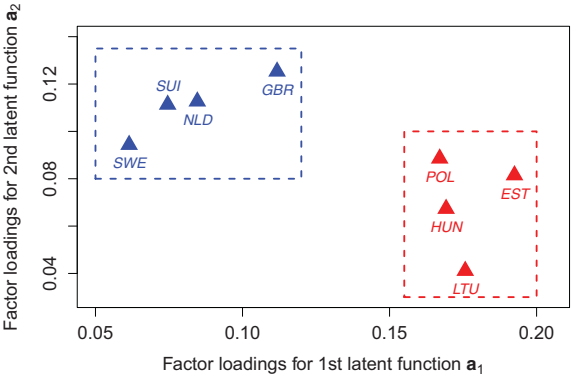


Figure 3. Factor loadings $a_{l,q}$ in the eight-population ICM with $Q=2$ in Table 1 – Case study II. The factor with the higher loading for each country is bolded in the Table on the left.

3.2 Incorporating latest data from other populations

In HMD, the reported data from different countries arrive 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 modelled. This implies that the most recent observations may be dropped for some countries. Of course, such recent data are 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.

To assess the value of information fusion and its link to population cross-correlation, we investigate the improvement in prediction in MOGP over SOGP. To do so, we set up a "notched" two-population training set where the foreign population has one more year of data and the assessment is based on one-year-ahead prediction for the domestic target population. Note that

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

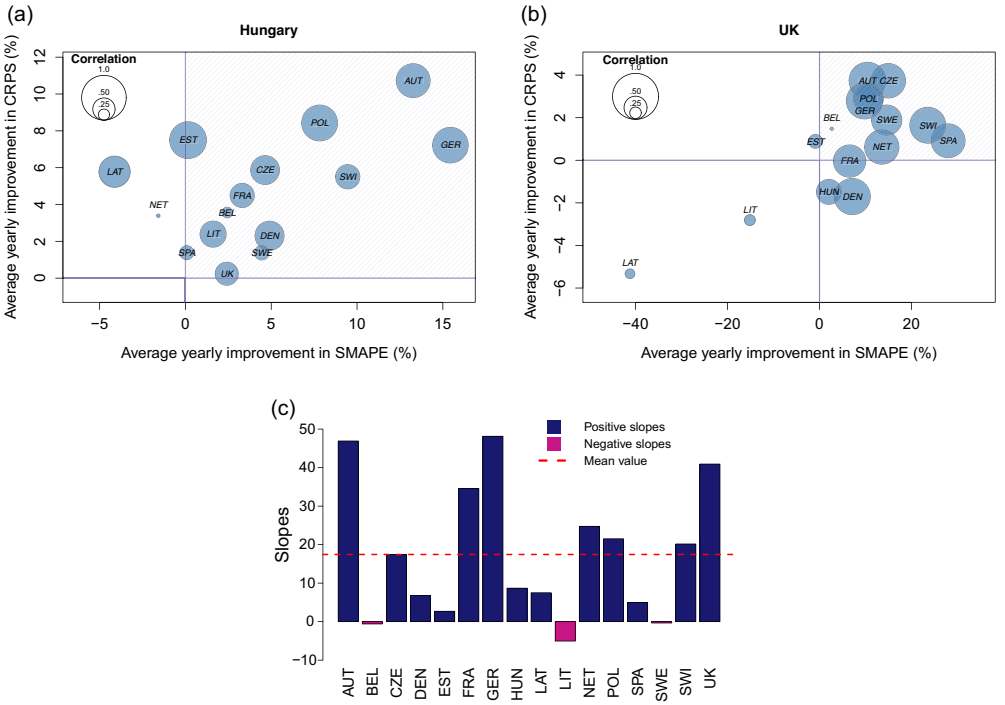


Figure 4. Prediction improvements for Hungarian and UK Males and the impact of cross-correlation on the improvement in SMAPE across 16 European national populations. (a) Hungary: $\sigma_l^2 = 2.083 \times 10^{-3}$. (b) UK: $\sigma_l^2 = 6.834 \times 10^{-4}$. (c) Linear slope of improvement in SMAPE versus correlation.

such "notch" extrapolation is not possible in the Lee–Carter framework that requires rectangular datasets. All ICM MOGP models take $L = 2$ and are fitted on Ages 70–84 in three different time frames: period 1990–2013/2014 for 2014 forecast, 1990–2014/2015 for 2015 forecast and 1990–2015/2016 for 2016 forecast. The comparator training datasets for SOGP models did not have mortality information in the calendar year of forecast (e.g. training on 1990–2013 for 2014 forecast). We report the resulting 3-year average percent improvement in CRPS and SMAPE between MOGP and SOGP. The positive average improvement is equivalent to MOGP models having smaller SMAPE and CRPS.

Figure 4 displays the improvement in SMAPE and CRPS of two-population MOGPs versus SOGP. We plot the results against the correlation r_{l_1,l_2} between the two populations modelled in MOGP and consider two randomly chosen target populations: Males in Hungary and Males in UK. Shaded regions indicate which MOGP models have both SMAPE and CRPS values less than the baseline. We observe that joint modelling generally yields higher improvement in Hungary compared to UK which is driven by the former's relatively smaller population which translates into larger σ_l and more opportunity for information fusion. Thus, for UK the single-population model is often competitive in its forecasting performance with a two-population MOGP.

The above discussion suggests that fusing highly correlated populations is better for predictive accuracy. To confirm this hypothesis, Figure 4(c) summarises the relationship between SMAPE improvement and correlation r_{l_1,l_2}. We used all 16 populations as targets and built 16×15 two-population MOGPs to record the resulting MOGP-SOGP improvements in SMAPE like in the right and middle panels. For each target population, we then fitted a linear regression model treating the 3-year average improvement in SMAPE as the dependent outcome and the correlation as the independent variable: SMAPE CHANGE $= b_0 + b_1 r_{l_1,l_2}$. Figure 4(c) displays the resulting slopes b_1 across the 16 populations. Positive b_1 implies that higher correlation leads to lower

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

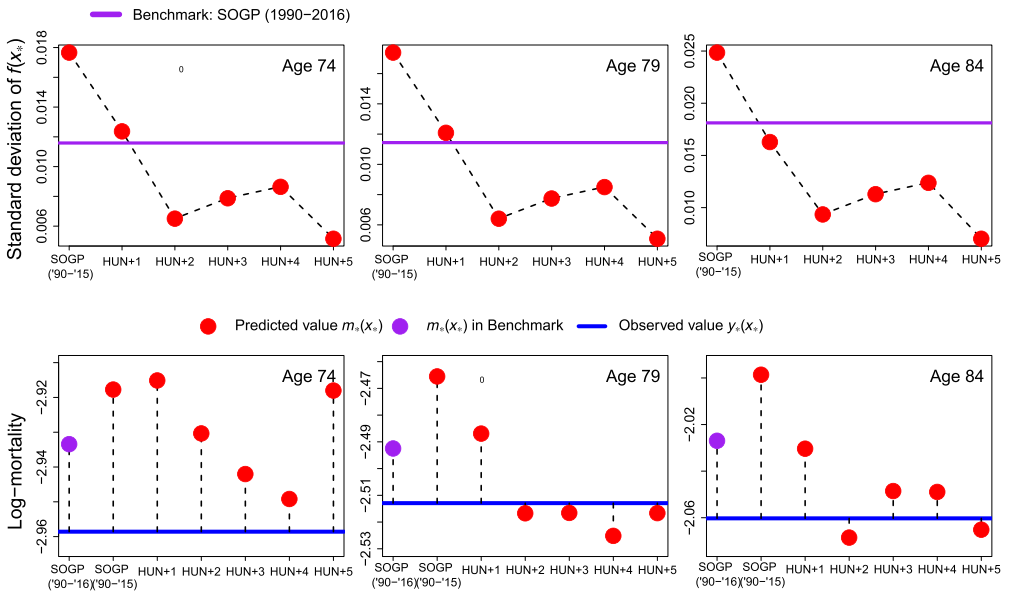


Figure 5. Comparison of prediction accuracy for 2016 log-mortality of Hungarian Males between different ICM-MOGP models with "notched" setup. Top row: standard deviation of $f(x_*)$; bottom row: distance between predicted mean $m_*(x_*)$ and the observed value $y_*(x_*)$.

SMAPE, i.e. higher predictive accuracy. The mean value of the b_1-slopes is highly positive and is around 20%. These empirical features suggest that one should indeed focus on aggregating *related* populations and discard unrelated ones. This is consistent with the results in Table 1 earlier. Most populations in Case Study II are less correlated with Hungary compared to Case Study I and as a result the SMAPE improvement in Case Study I is higher (10.75%) compared to Case Study I (−2.77%).

Moving beyond two populations, in Figure 5, we illustrate predictive gains for Hungarian Males due to incorporating most recent foreign data. This complements Table 3 that considered an isotropic dataset, with a notched setup instead. Our benchmark is a Hungary Males SOGP model fitted on 1990–2016. We then drop 2016 Hungary observations, but augment with 1990–2016 data from *other* countries and perform 1-year-out extrapolation to forecast 2016 Hungary mortality. These models are labelled as "HUN+1", "HUN+2", etc., to indicate the number of foreign populations considered. The top panels of Figure 5 visualise increasing forecast credibility, namely lower $s_*(x_*)$, for Hungary as more and more correlated data are added. In fact, we see that a MOGP 1-year-out prediction with 3+ populations is *more* credible than direct smoothing of realised 2016 Hungary experience. As expected, we observe that credibility gains flatten out as L continues to grow and available information is saturated. The bottom panels of Figure 5 display the prediction errors $m_*(x_*) - y_*(x_*)$ relative to realised 2016 Hungary experience. Again, we see that higher L tends to generate less bias in prediction, confirm the earlier SMAPE analysis for isotropic case studies. Due to the strong observation noise, the pattern for a specific cell x_* can be erratic, although in nearly all cases, MOGP easily beats out the plain SOGP. To conclude, borrowing latest information from nearby highly correlated populations is essentially as good as having the latest domestic data, and is significantly better than just using the available domestic data.

Remark: In Figure 5 (and earlier in Table 3), we add populations based on their correlation to the target population, i.e. we pool through estimated θ_{l_1,l_2}. In Appendix A, we discuss a simpler alternative based on comparing mean functions that capture historical mortality trends and then running a hierarchical clustering method.

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

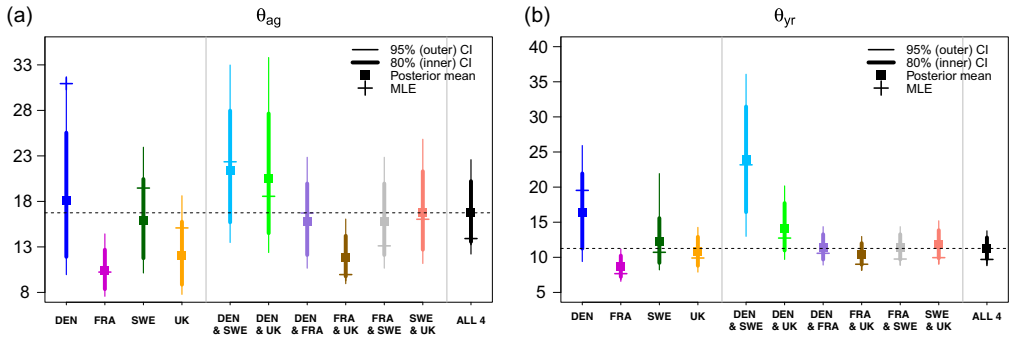


Figure 6. `Stan` MCMC posteriors of the length-scales θ for Age and Year across populations and joint models with different groupings. The $+$'s indicate the respective MLE estimates from a `kergp` model. The dashed lines indicate the MCMC MAP estimate from the four-population full-MOGP model. (a) Length-scale in the Age dimension. (b) Length-scale in the Year dimension.

4. Further Features of MOGP Models

4.1 Improved hyperparameter estimation

To illustrate the commonality in mortality experience in related populations, we perform Bayesian GP on four developed Western European countries: Sweden, Denmark, France and UK. Figure 6 shows the inferences of the length-scales for Age and Year along with MLE estimations when fitting SOGP models for each population versus jointly modelling them as groups of two, or jointly as all four together. The figures visualise 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 (Figure 6(a)). However, once we pair Denmark with either Sweden, UK or France (Figure 6(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 four 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 6, the four-population MAP estimates of the length-scales (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. We also remark that the MLE estimates fall within the 95% posterior credible intervals from the `Stan` model indicating that Bayesian inference works properly.

This also highlights the ability of joint models to better estimate the hyperparameters by utilising multiple datasets. It is known that GPs might have difficulties in estimating length-scales, for example due to the likelihood function (15) being highly multi-modal, or conversely very flat around its maxima. Providing more data is one remedy. As discussed in Huynh *et al.* (2020), some SOGP will over- or under-smooth data while pooling data across multiple populations achieves shrinkage towards the global hyperparameter mean and provides a better fit.

4.2 Coherent mortality forecasts

Fitting GP models for individual populations tends to generate divergent long-term forecasts that are inconsistent with historical patterns. MOGP models do not have this limitation and maintain the historical characteristics observed in the data into the future. Namely, in MOGP 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 modeller. For the linear mean function in (14), the population coefficients $\beta_{pop,l}$ serve this purpose and represent the

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

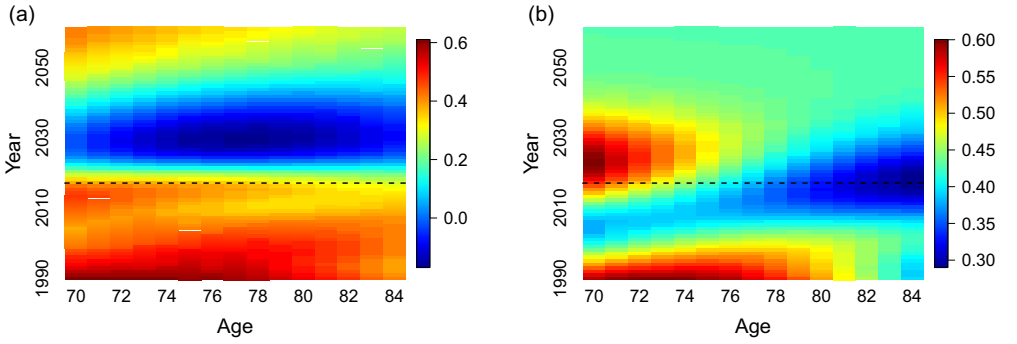


Figure 7. Forecasted mean difference between Danish Male and Female mortality over 1990–2060. Training set was Ages 70–84 and Years 1990–2016 (edge of training set indicated by the dashed lines). (a) Male/Female difference in log-mortality in Denmark using individual (SOGP) models. (b) Male/Female difference in log-mortality in Denmark using a joint (full-rank MOGP) model.

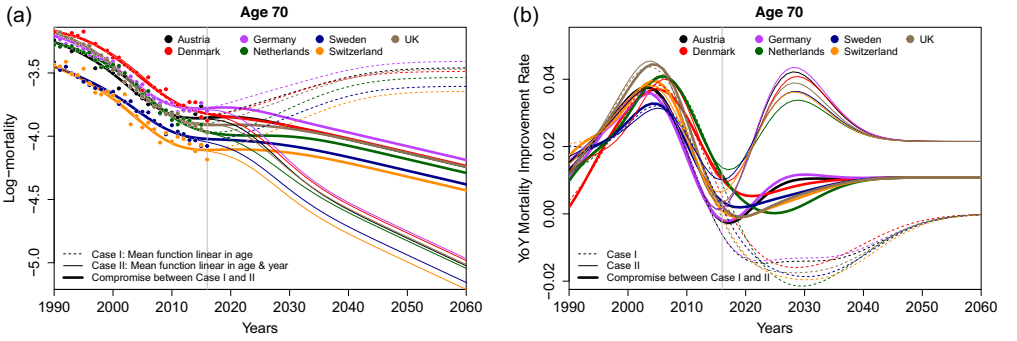


Figure 8. Long-term mortality forecasting over years 1990–2060. The models are based on ICM kernel with $Q = 3$ and are trained using Ages 70–84 and Years 1990–2016 over seven Male populations: Austria, Denmark, Germany, Netherlands, Sweden, Switzerland and UK. Dashed lines indicate the boundary between in-sample and out-of-sample forecasts. (a) Predicted log-mortality rate. (b) Annual mortality improvement factors.

long-term spread between same-Age log-mortality rates. To illustrate the above, consider mortality differences due to gender. Women outlive men by 7 years on average in developed countries (United Nations, 2011). Modelling mortality for each gender separately fails to take into account this interdependence and tends to result in divergent and implausible long-run forecasts even if the same fitting procedure is applied. The heatmaps in Figure 7 display the projected Male–Female differences in log-mortality for Denmark; single-population models on the left imply that as early as 2030, Males will have lower mortality than females. In contrast, the MOGP forecast in the right panel is coherent: Females are projected to maintain higher longevity and historical patterns slow dissipate over time to the long-term gap of $\beta_{pop,2} - \beta_{pop,1} = 41.5\%$ between same-age Male and Female mortality.

Figure 8 shows the log-mortality and annual mortality improvement rates for Males aged 70 across seven European populations (indicated by colours), in the period from 1990 to 2060. We build a seven-population MOGP based on three different scenarios for $m(x)$:

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

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

2. Long-term mortality improvement based on a historical pattern (thin solid curves). This is encapsulated via $m(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \beta_1^{yr} x_{yr}^n + \sum_{l=2}^{L} \beta_{pop,l} x_{pop,l}^n$. In the long-run $\partial m_{back}^{GP}(.;yr) \to \beta_1^{yr}$ (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 $m(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \beta_1^{yr} x_{yr}^n + \sum_{l=2}^{L} \beta_{pop,l} \left(x_{pop,l}^n\right)$, but this time the β_1^{yr} coefficient is picked by the modeller and for illustrative purposes fixed at 1% to reflect recent slowdown in global MI. 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 & Tickle, 2008).

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 length-scale $\theta_{yr} \simeq 10$) 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 Denmark surpassing Germany's longevity) due to 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.

Remark: As noted throughout our analysis, the GP framework is based on exploiting the spatial correlation between data points. This data-driven nature of our model makes it not well-suited for long-range forecasts. The long-term prediction from the GP converges to its prior mean as the correlation between the historical training data and the future data points that are decades apart declines to zero. This de-correlation also leads to very high forecast uncertainty. With this in mind, the primary usage of our models is for "nowcasting" and medium-term (less than 10 years) uncertainty quantification of joint longevity experiences. Accurate short-term projections have been important for instance in the context of quantifying COVID-19 excess deaths, see https://nhanhuynh46.github.io/MOGPTutorials/SOGP_Covid19.html.

5. Conclusion

We have developed and investigated stochastic multi-population mortality models based on MOGP regression. In our approach, cross-population dependence is captured via spatial correlation that overlays the Age–Year structure. This yields a unified approach for any number of populations L; moreover, the proposed coregionalisation kernels allow to leverage the Kronecker structure and incorporate dimension reduction for the underlying cross-population factors. Our analysis of HMD data suggests that the MOGP approach is well-suited to selectively fuse mortality experience from similar datasets, where similarity can be interpreted through the spatial GP correlations r_{l_1,l_2}. On the one hand, we find that pooling disparate populations can be counter-productive (since MOGP relies on the assumption of homogeneous Age–Year covariance pattern); on the other hand, pooling can indeed yield significant improvement in predictive accuracy, especially in smaller populations with low credibility.

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 modelling infrastructure as the presented approach becomes computationally expensive for $L \gg 10$ populations (more than $N \gg 5,000$ total cells). There is currently a very active and ongoing progress on large-scale GP models especially for gridded data like in HMD, see, e.g., Flaxman *et al.* (2015). A different avenue of future research would be to systematically explore the best spatial covariance structures, as encapsulated by the kernel function $\tilde{C}(x, x')$. In this paper, we focused on only using the SE kernel

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

and standard Age- and Year-effects. It is feasible to consider further dependence formats, e.g. Birth Cohort effect (see Appendix C and Figure C.1 below), and other kernel families, such as the Matérn (Ludkovski *et al.*, 2018). A third direction would be to revisit the observation variance assumption via GLM (generalised linear model) GP formulations.

Acknowledgement. We thank the anonymous referees and the associate editor for many useful suggestions. We also thank Howard Zail for several informative discussions. Ludkovski is partially supported by NSF DMS-1821240.

References

Alvarez, **M.A.**, **Rosasco**, **L.** & **Lawrence**, **N.D.** (2011). Kernels for vector-valued functions: a review.

Antonio, **K.**, **Devriendt**, **S.**, **de Boer**, **W.**, **de Vries**, **R.**, **De Waegenaere**, **A.**, **Kan**, **H.-K.**, **Kromme**, **E.**, **Ouburg**, **W.**, **Schulteis**, **T.**, **Slagter**, **E. et al.** (2017). Producing the Dutch and Belgian mortality projections: a stochastic multi-population standard. *European Actuarial Journal*, **7**(2), 297–336.

Armstrong, **J.S.** & **Collopy**, **F.** (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting*, **8**(1), 69–80.

Betancourt, **M.** (2017). Robust Gaussian processes in Stan. Available online at the address https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html.

Boe, **C.**, **Winant**, **C.**, **Riffe**, **T.**, **Barbieri**, **M.**, **Wilmoth**, **J.R.**, **Jasilionis**, **D.**, **Grigoriev**, **P.**, **Jdanov**, **D.**, **Shkolnikov**, **V.M.** & **Glei**, **D.** (2015). Data resource profile: The Human Mortality Database (HMD). *International Journal of Epidemiology*, **44** (5), 1549–1556.

Bonilla, **E.V.**, **Chai**, **K.M.** & **Williams**, **C.** (2008). Multi-task Gaussian process prediction. In **J.C. Platt**, **D. Koller**, **Y. Singer** & **S.T. Roweis** (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 153–160). Curran Associates, Inc.

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

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

Carpenter, **B.**, **Gelman**, **A.**, **Hoffman**, **M.D.**, **Lee**, **D.**, **Goodrich**, **B.**, **Betancourt**, **M.**, **Brubaker**, **M.**, **Guo**, **J.**, **Li**, **P.** & **Riddell**, **A.** (2017). Stan: a probabilistic programming language. *Journal of Statistical Software*, **76**(1).

Carracedo, **P.**, **Debón**, **A.**, **Iftimi**, **A.** & **Montes**, **F.** (2018). Detecting spatio-temporal mortality clusters of European countries by sex and age. *International Journal for Equity in Health*, **17** (1), 38.

Caruana, **R.** (1997). Multitask learning. *Machine Learning*, **28**(1), 41–75. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

Chen, **H.**, **MacMinn**, **R.** & **Sun**, **T.** (2015). Multi-population mortality models: a factor copula approach. *Insurance: Mathematics and Economics*, **63**, 135–146.

Chiles, **J.-P.** & **Delfiner**, **P.** (1999). *Geostatistics: Modeling Spatial Uncertainty*. Wiley. ISBN 0471083151 9780471083153.

Christiansen, **M.C.**, **Spodarev**, **E.** & **Unseld**, **V.** (2015). Differences in European mortality rates: a geometric approach on the age–period plane. *ASTIN Bulletin: The Journal of the IAA*, **45** (3), 477–502.

Chu, **W.** & **Ghahramani**, **Z.** (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, **6**, 1019–1041.

D'Amato, **V.**, **Haberman**, **S.**, **Piscopo**, **G.**, **Russolillo**, **M.** & **Trapani**, **L.** (2016). Multiple mortality modeling in Poisson Lee–Carter framework. *Communications in Statistics-Theory and Methods*, **45** (6), 1723–1732.

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

Delwarde, **A.**, **Denuit**, **M.**, **Guillén**, **M.** & Vidiella-i Anguera, A. (2006). Application of the Poisson log-bilinear projection model to the G5 mortality experience. *Belgian Actuarial Bulletin*, **6**(1), 54–68.

Deville, **Y.**, **Ginsbourger**, **D.** & **Roustant**, **O.** (2019). Contributors: Nicolas Durrande. *kergp: Gaussian Process Laboratory*. Available online at the address https://CRAN.R-project.org/package=kergp. R package version 0.5.0.

Dong, **Y.**, **Huang**, **F.**, **Yu**, **H.** & **Haberman**, **S.** (2020). Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal*, 2020(8), 754–775.

Duvenaud, **D.K.** (2014). *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge.

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

Flaxman, **S.**, **Gelman**, **A.**, **Neill**, **D.**, **Smola**, **A.**, **Vehtari**, **A.** & **Wilson**, **A.G.** (2015). Fast hierarchical Gaussian processes, technical report, Preprint at http://sethrf.com/files/fast-hierarchical-GPs.pdf.

Garrido-Merchán, **E.C.** & **Hernández-Lobato**, **D.** (2018). Dealing with categorical and integer-valued variables in Bayesian optimization with Gaussian processes. CoRR, abs/1805.03463.

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

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

HMD. (2018). The Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available online at the address www.mortality.org.

Hoef, **J.M.V.** & **Barry**, **R.P.** (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, **69** (2), 275–294. ISSN 0378-3758.

Huynh, **N.**, **Ludkovski**, **M.** & **Zail**, **H.** (2020). Multi-population longevity models: a spatial random field approach. In *Proceedings of the Society of Actuaries 2020 Living to 100 Symposium*.

Hyndman, **R.J.**, **Booth**, **H.** & **Yasmeen**, **F.** (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, **50**(1), 261–283.

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

Kleinow, **T.** & **Cairns**, **A.J.G.** (2013). Mortality and smoking prevalence: an empirical investigation in ten developed countries. *British Actuarial Journal*, **18**(2), 452–466.

Letham, **B.** & **Bakshy**, **E.** (2019). Bayesian optimization for policy search via online-offline experimentation.

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

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

Li, **N.** & **Lee**, **R.** (2005). Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography*, **42**(3), 575–594.

Lu, **Q.**, **Hanewald**, **K.** & **Wang**, **X.** (2019). Bayesian hierarchical multi-population mortality modelling for china's provinces. SSRN. Available online at the address doi: http://dx.doi.org/10.2139/ssrn.3494491.

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

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

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

Qin, **C.** & **Jevtic**, **P.** (2016). Multi-population mortality modelling with levy processes. SSRN. Available online at the address doi: http://dx.doi.org/10.2139/ssrn.2813678.

Raftery, **A.E.**, **Li**, **N.**, **Ševčková**, **H.**, **Gerland**, **P.** & **Heilig**, **G.K.** (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109, 13915–13921.

Rasmussen, **C.E.** & **Williams**, **C.K.I.** (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Roustant, **O.**, **Ginsbourger**, **D.** & **Deville**, **Y.** (2012). 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.

Roustant, **O.**, **Padonou**, **E.**, **Deville**, **Y.**, **Clément**, **A.**, **Perrin**, **G.**, **Giorla**, **J.** & **Wynn**, **H.P.** (2018). Group kernels for Gaussian process metamodels with categorical inputs. Working Paper or preprint, July 2018. Available online at the address https://hal.archives-ouvertes.fr/hal-01702607.

Shang, **H.L.** (2016). 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.

Skolidis, **G.** & **Sanguinetti**, **G.** (2011). Bayesian multitask classification with Gaussian process priors. *IEEE Transactions on Neural Networks*, **22** (12), 2011–2021. ISSN 1941-0093.

Tsai, **C.C.-L.** & **Wu**, **A.D.** (2020). Incorporating hierarchical credibility theory into modelling of multi-country mortality rates. *Insurance: Mathematics and Economics*, **91**, 37–54. ISSN 0167-6687.

Tsai, **C.C.-L.** & **Zhang**, **Y.** (2019). A multi-dimensional bhlmann credibility approach to modeling multi-population mortality rates. *Scandinavian Actuarial Journal*, **2019**(5), 406–431.

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

Wang, **C.-W.**, **Yang**, **S.S.** & **Huang**, **H.-C.** (2015). Modeling multi-country mortality dependence and its application in pricing survivor index swapsÂ"–a dynamic copula approach. *Insurance: Mathematics and Economics*, **63**, 30–39.

Wang, **P.**, **Pantelous**, **A.A.** & **Vahid**, **F.** (2020). Multi-population mortality projection: the augmented common factor model with structural breaks. SSRN.

Wiśniowski, **A.**, **Smith**, **P.W.F.**, **Bijak**, **J.**, **Raymer**, **J.** & **Forster**, **J.J.** (2015). Bayesian population forecasting: extending the Lee-Carter method. *Demography*, **52**(3), 1035–1059.

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

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Appendix A. Clustering by Mortality Trends

In Section 3.1, we constructed multi-population MOGPs by first generating a large set of two-population models and then utilising the respective correlations r_{l_1,l_2} to select the datasets that are most correlated to the target population. We also investigated a simpler alternative for deciding which populations to pool based on the similarity in their mortality trends. This approach does not require construction of any preliminary MOGPs and instead only looks at the GP mean function $m(\cdot)$. Namely, we first estimate the shape of mortality for each population l via a linear mean function: $m_l(x) = \beta_{0,l} + \beta^{ag}_{1,l} x_{ag} + \beta^{yr}_{1,l} x_{ag}$. We then calculate the root-integrated squared distance between $m_{l_1}(\cdot)$, $m_{l_2}(\cdot)$ based on the given test set $x_{ag} \in \{50, \ldots, 84\}$ and $x_{yr} \in \{1990, ..., 2012\}$:

$$D_{l_1,l_2} = \sqrt{\int_{1990}^{2012} \int_{50}^{84} \left[(\beta_{0,l_1} - \beta_{0,l_2}) + (\beta^{ag}_{1,l_1} - \beta^{ag}_{1,l_2}) x_{ag} + (\beta^{yr}_{1,l_1} - \beta^{yr}_{1,l_2}) x_{yr} \right]^2 dx_{ag} dx_{yr}}. \quad \text{(A.1)}$$

The above metric is employed within a hierarchical clustering method based on a specified dissimilarity measure. Figure A.1 displays the dendrograms extracted from hierarchical clustering of 32 populations (16 countries × {Male, Female}) via two different measures of dissimilarity: single linkage and complete linkage. We note that the resulting clusters naturally tend to separate males and females, reflecting the latter's lower mortality. We also observe a strong geographic influence, so that neighbouring countries with similar demographics are indeed clustered together. The dendrogram could be used agglomeratively to build a MOGP. For example, using Hungary as the target population, Figure A.1 suggests first adding Estonia, then Lithuania, Latvia, etc. We find that this method is often as efficient as clustering by cross-population correlation, although going up the linkages often calls for fusing of two clusters, i.e. it is not designed for increasing L 1-by-1.

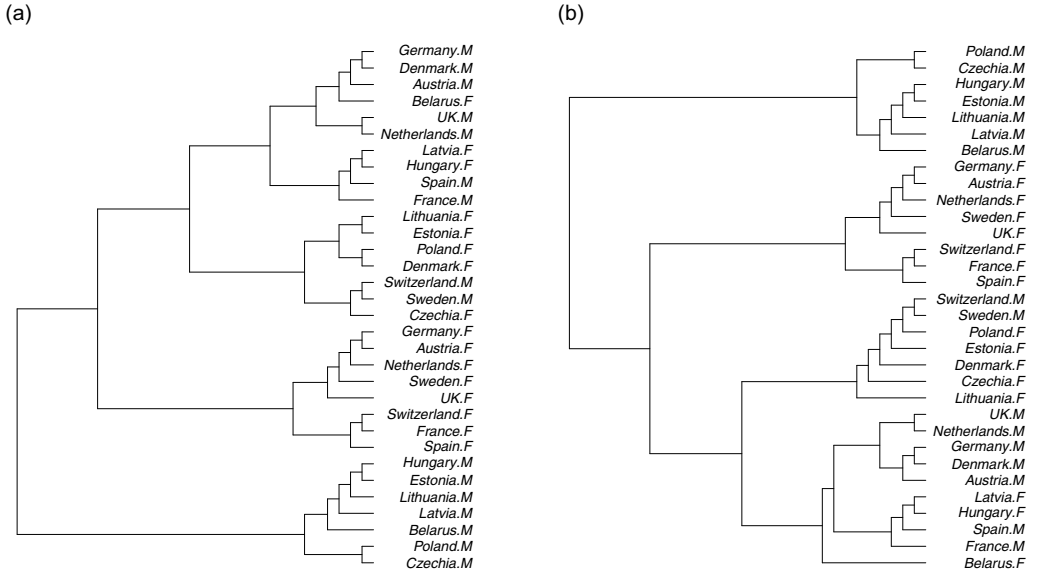


Figure A.1. Dendrograms from hierarchical clustering of 32 HMD populations using the *D*-metric in (A.1). (a) Complete linkage. (b) Single linkage.

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Appendix B. Fitted MOGP Hyperparameters Across Full-Rank and ICM Kernels and MLE/Bayesian Methods for the DEN-FRA-SWE-GBR Case Study

Cross-population correlation matrices. In the full-rank kernel (7), the correlation coefficient is $r_{l_1,l_2} := \exp(-\theta_{l_1,l_2})$. In the ICM model, the cross-covariance is $B = AA^T$ with the diagonals $B_{l,l}$ interpreted as the process variance of f_l, from which we can similarly extract r_{l_1,l_2}'s. Using {Denmark, France, Sweden, UK} $\equiv \{1, 2, 3, 4\}$ we find that

$$\begin{bmatrix} r_{21} & & \\ r_{31} & r_{32} & \\ r_{41} & r_{42} & r_{43} \end{bmatrix} = \begin{bmatrix} 0.82 & & \\ 0.83 & 0.90 & \\ 0.93 & 0.90 & 0.84 \end{bmatrix} \text{ for full rank; and}$$

$$= \begin{bmatrix} 0.75 & & \\ 0.73 & 0.84 & \\ 0.96 & 0.98 & 0.89 \end{bmatrix} \text{ for ICM.}$$

Table B.1. Hyperparameter estimates based on maximum likelihood (`kergp`) 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. Denmark used as baseline population in the mean function

Parameters	`kergp` MLE	`Stan`		
		MAP	MCMC Mean	MCMC 95% Posterior CI
β_0	−10.4610	−10.0220	−10.5337	(−12.0847, −9.1274)
β_1^{ag}	0.0996	0.0958	0.0967	(0.0847, 0.1085)
β_{FRA}	−0.0922	−0.0685	0.1239	(−0.2438, 0.5827)
β_{SWE}	−0.1076	−0.0592	−0.0060	(−0.3596, 0.3844)
β_{GBR}	0.0062	0.000	0.1122	(−0.2252, 0.4961)
θ_{ag}	13.3849	12.1915	17.4166	(12.0294, 24.0641)
θ_{yr}	8.8549	9.2694	11.3858	(8.2536, 13.3009)
$\theta_{DEN,\,FRA}$	0.1956	0.3773	0.8269	(0.1544, 2.9089)
$\theta_{DEN,\,SWE}$	0.1804	0.2725	0.5094	(0.0889, 1.8891)
$\theta_{DEN,\,GBR}$	0.0772	0.0799	0.1579	(0.0286, 0.5473)
$\theta_{FRA,\,SWE}$	0.1030	0.1943	0.3658	(0.0797, 1.0949)
$\theta_{FRA,\,GBR}$	0.1095	0.1445	0.1439	(0.0383, 0.3917)
$\theta_{SWE,\,GBR}$	0.1681	0.1801	0.6132	(0.0530, 2.6660)
η^2	0.0395	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_{GBR}	6.887×10^{-4}	6.849×10^{-4}	7.001×10^{-4}	$(5.985 \times 10^{-4}, 8.165 \times 10^{-4})$

Appendix C. Alternative Kernel Families

In this Appendix, we illustrate selecting alternative kernel families for the GP model as discussed in Section 2.6. Figure C.1(a) and (b) compares employing a Matérn-5/2 kernel instead of the squared exponential one. Namely, we replace the kernel definition with

$$C(x, x') = \eta^2 \prod_{k \in \{ag, yr\}} \left(1 + \frac{\sqrt{5}}{\theta_k}|x_k - x'_k| + \frac{5}{3\theta_k^2}(x_k - x'_k)^2\right) e^{-\frac{\sqrt{5}}{\theta_k}|x_k - x'_k|}$$

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

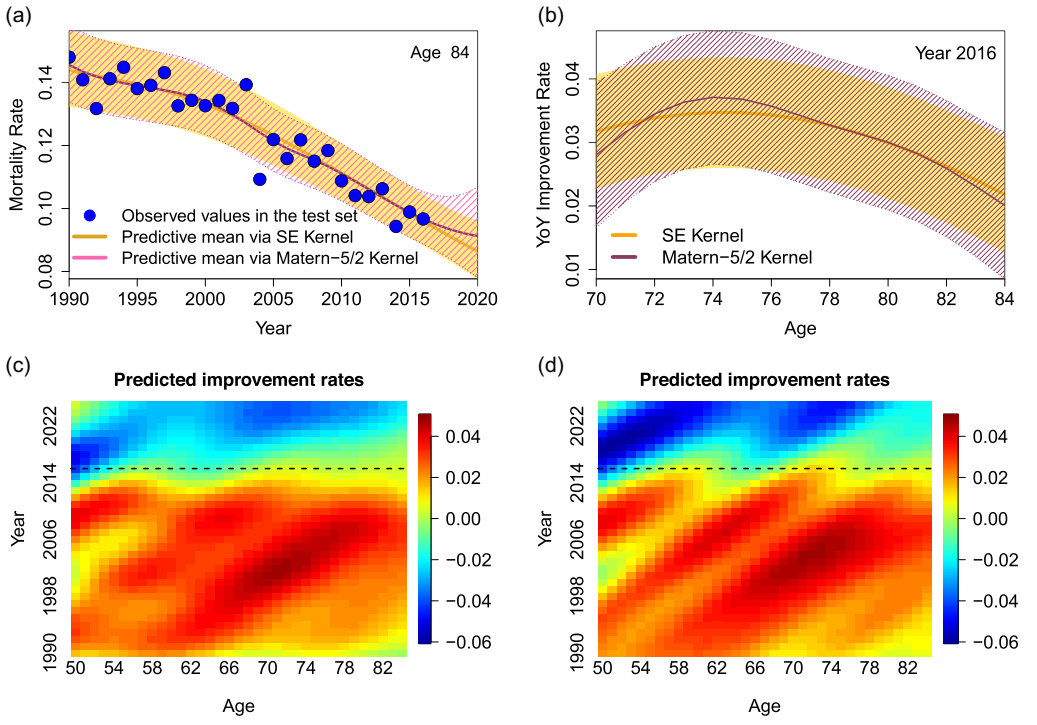


Figure C.1. Illustration I: Comparing a GP model with a Matérn-5/2 covariance kernel vis-a-vis one with a squared-exponential kernel. We show mortality rates and improvement rates for Danish Males. Both models are fitted via the ICM approach and use four populations from Denmark, France, Sweden and UK; Males, Ages 70–84 and Years 1990–2016. **Illustration II:** Heatmap of the predicted mortality improvement rates using a squared exponential kernel without and with a Birth–Cohort term (17) for UK Males. The single-population model is fitted on Ages 50–84 and Years 1990–2016. Effect of alternative covariance kernel specifications on the GP model output. (a) Mortality rates for Danish Males. (b) YoY improvement rates for Danish Males. (c) Base case (no Cohort). (d) With Birth-Cohort.

The results are illustrated via the smoothed mortality rates and the respective improvement rates for Danish Males. We observe that the Matérn-5/2-based fit hews closer to the historical data which leads to more "rough" mortality improvement factors. Otherwise, the differences are mild.

Table C.2 illustrates the impact of implementing a composite kernel, namely the sum of a linear kernel and a SE kernel in (16), see section 2.6. We again consider Male mortality in Denmark within a single-population setup. The selected composite kernel is functionally equivalent to an SE kernel with a linear trend function (Duvenaud, 2014), although the obtained results differ slightly since the optimised marginal likelihood functions are not algebraically the same. We do observe essentially the same results (they are visually identical which is the reason for not including a plot). Note that in Table C.2 the hyperparameters associated with the linear kernel, α_0^2 and α_{ag}^2, can be thought as the priors of the intercept and the slope in Age, respectively. In contrast, with a parametric mean function, β_0 is interpreted as the intercept and β_1 as the linear trend in Age.

Section 2.6 also briefly discussed the potential to incorporate Year-of-Birth term into the covariance structure to capture cohort effects, cf. (17). Cohort effects manifest themselves in diagonal features on an Age–Year heatmap, see Figure C.1(c)–(d) where we illustrate with data for UK Males Ages 50–84 and Years 1990–2016. In the figure, we display model-predicted mortality improvement rates covering the in-sample cells, as well as forecasting out to year 2025. The bottom left panel (c) uses the base SE kernel without a Birth–Year term; one can clearly see several

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

Table C.1. MOGP with rank $Q = 3$ ICM kernel for Males in Denmark, France, Sweden and UK. The training dataset contains Ages 70–84 and Years 1990–2012

Mean function			Length-scales	Cross-covariance	$a_{l,1}$	$a_{l,2}$	$a_{l,3}$
β_0	β_{ag}	β_{FRA}	θ_{ag}	DEN: $l = 1$	0.0435	0.1687	0.1068
−11.4073	0.1120	0.0410	15.4199	FRA: $l = 2$	0.2199	0.1619	0.0491
–	β_{SWE}	β_{GBR}	θ_{yr}	SWE: $l = 3$	0.1552	0.0562	0.1683
–	−0.0348	0.0520	13.4019	GBR: $l = 4$	0.1451	0.1636	0.1640

Table C.2. Comparison between the hyperparameters from fitting a fused kernel vis-a-vis a composite Linear+SE kernel (16) for a SOGP model for Danish Males, Ages 70–84 and Years 1990–2016

	Hyperparameter	SE kernel	Linear + SE kernel
Mean function	β_0	−10.1560	0
	β_{ag}	0.0951	–
Covariance function	α_0^2	–	36.6177
	α_{ag}^2	–	0.0082
	θ_{ag}	12.5430	12.1455
	θ_{yr}	15.1672	18.2727
	η^2	0.0598	0.0862
	σ^2	$1.589e-03$	$1.592e-03$

Table C.3. GP hyperparameters for fitting a covariance kernel with Birth–Cohort term (17) for a SOGP model for UK Males, Ages 50–84 and Years 1990–2016

Hyperparameter		Base case SOGP without cohort	SOGP with Birth-Cohort term
Mean function	β_0	−10.3695	−10.1886
	β_{ag}	0.0976	0.0961
Covariance function	θ_{ag}	6.5553	58.1212
	θ_{yr}	5.5229	11.5091
	θ_{bc}	–	6.5911
	η^2	0.0239	0.0324
	σ^2	$5.421e-04$	$5.260e-04$

diagonal patterns for birth cohorts circa 1930, c. 1957 (both of which enjoyed larger than average MI) and c. 1965 (which is experiencing negative MI).

To explicitly capture these features, in the bottom-right panel (d) we construct the same heatmap but now with covariance kernel $C(x, x')$ of the form (17). While the in-sample patterns remain the same (as they should, being primarily data-driven), we can spot differences in terms of out-of-sample forecasts. The Cohort-GP model makes the historical cohort trends persist into the future; for example it makes the longevity of the post-1945 cohort to continue to be materially worse than the World War II generation (extreme top-right corner), while the no-Cohort model expects longevity deterioration in the 2020s across all cohorts.

Adding a Birth-year term to $C(x, x')$ changes all the GP hyperparameters, especially the length-scales, cf. Table C.3. Since (17) has a product of three terms rather than two as before, to maintain a similar correlation between a given x, x', each length-scale is attenuated. Thus, θ_{yr} more than

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press

doubles from ≈ 5.5 to ≈ 11.5, while θ_{ag} increases even more dramatically from ≈ 6.5 to ≈ 58 when a Birth-Cohort term is included. The estimated Birth–Cohort length-scale, $\theta_{bc} \approx 6.6$, indicates that the typical "generation" spans about 7 birth–years, e.g. 1939–1945, 1965–1972. As expected, the change should have minimal impact on the GP mean function or the variance parameters η^2, σ^2.

Cite this article: Huynh N and Ludkovski M (2021). Multi-output Gaussian processes for multi-population longevity modelling. *Annals of Actuarial Science* **15**, 318–345. https://doi.org/10.1017/S1748499521000142

https://doi.org/10.1017/S1748499521000142 Published online by Cambridge University Press