Annals of Actuarial Science (2024), **18**, pp. 51–77
doi:10.1017/S1748499523000118

ORIGINAL RESEARCH PAPER

Joint models for cause-of-death mortality in multiple populations

Nhan Huynh and Mike Ludkovski

Department of Statistics and Applied Probability, University of California at Santa Barbara, Santa Barbara, CA 93106, USA
Corresponding author: Mike Ludkovski; Email: ludkovski@pstat.ucsb.edu

(Received 12 November 2021; revised 01 April 2023; accepted 24 April 2023; first published online 18 May 2023)

Abstract

We investigate jointly modelling age–year-specific rates of various causes of death in a multinational setting. We apply multi-output Gaussian processes (MOGPs), a spatial machine learning method, to smooth and extrapolate multiple cause-of-death mortality rates across several countries and both genders. To maintain flexibility and scalability, we investigate MOGPs with Kronecker-structured kernels and latent factors. In particular, we develop a custom multi-level MOGP that leverages the gridded structure of mortality tables to efficiently capture heterogeneity and dependence across different factor inputs. Results are illustrated with datasets from the Human Cause-of-Death Database (HCD). We discuss a case study involving cancer variations in three European nations and a US-based study that considers eight top-level causes and includes comparison to all-cause analysis. Our models provide insights into the commonality of cause-specific mortality trends and demonstrate the opportunities for respective data fusion.

Keywords: Cause-of-death modelling; Gaussian process models; Multiple populations; Multiple causes of death

1. Background and Motivation

In-depth modelling of the evolution of human mortality necessitates analysis of the prevalent causes of death. This is doubly so for making mortality forecasts into the future across different age groups, populations and genders. In this article, we develop a methodology for probabilistic forecasting of cause-specific mortality in a multi-population (primarily interpreted as a multinational) context. Thus, we simultaneously fit multiple cause-specific longevity surfaces via a spatio-temporal model that accounts for the complex dependencies across causes and countries and across the age–year dimensions.

While there have been many works on modelling mortality across several populations (Dong *et al*., 2020; Enchev *et al*., 2017; Guibert *et al*., 2019; Hyndman *et al*., 2013; Kleinow, 2015; Li & Lu, 2017; Tsai & Zhang, 2019), as well as an active literature on cause-of-death mortality, there are very few that do both simultaneously. As we detail below, there are many natural reasons for building such a joint model, and this gap is arguably driven by the underlying “Big Data” methodological challenge. Indeed, with dozens of mortality datasets that are indexed by countries, causes of death, genders, etc., developing a scalable approach is daunting. We demonstrate that this issue may be overcome by adapting machine learning approaches, specifically techniques from multi-task learning (Bonilla *et al*., 2008; Caruana, 1997; Letham & Bakshy, 2019; Williams *et al*., 2009). To this end, we employ multi-output Gaussian processes (MOGPs) combined with linear coregionalisation. GPs are a kernel-based data-driven regression framework that translates mortality modeling into smoothing and extrapolating an input–output response surface based

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

on noisy observations. It yields a full uncertainty quantification for mortality rates and mortality improvement factors. Coregionalisation is a dimension reduction technique that enables efficiently handling many correlated outputs.

This work is a continuation of our series of articles Ludkovski *et al.* (2018), Huynh *et al.* (2020) and Huynh & Ludkovski (2021) that discussed the application of GPs to model all-cause mortality in the single-population and multi-population contexts, respectively. Unlike all-cause mortality in different geographic regions, which tends to exhibit strong correlation and long-term coherence, different causes have less commonality and thus require a more flexible structure for the respective cross-dependence. Moreover, joint analysis of 10+ mortality surfaces brings computational scalability challenges of potentially hundreds of model parameters to calibrate. Thus, in order to carry out cause-specific mortality analysis, we make methodological innovations along two directions. First, we compare two distinct versions of the MOGP that implement dimension reduction by fusing outputs through linear combinations of latent functions: the semiparametric latent factor model (SLFM) and the intrinsic coregionalisation model (ICM). ICM assumes a fixed spatial kernel in age–year, while SLFM does not. This distinction corresponds to different assumptions about the structural commonality in the modeled mortality surfaces. We conduct sensitivity analysis to assess these two choices for the tasks of in-sample fitting and of forecasting cause-specific mortality. Second, we implement a multi-level MOGP-ICM model that separates latent factors across the different types (countries, causes, genders, etc.) of categorical inputs describing the populations. The separability assumption in the joint covariance kernel yields a product-type cross-population correlation structure, providing insights about the relationships between the mortality improvement trends.

Our models are driven by the scalability issue which has been a critical obstacle to analyse *in bulk* the large-volume mortality datasets that have become available recently. Thus, to cope with many mortality surfaces, we leverage the twin pillars of multi-output models (Teh *et al.*, 2005; Letham & Bakshy, 2019; Williams *et al.*, 2009) and the structured Kronecker covariance that mitigates the typical cubic computational complexity of GPs (Flaxman *et al.*, 2015; Gilboa *et al.*, 2015; Saatçi 2011; Zhe *et al.*, 2019).

Decomposing all-cause mortality rates leads to reduced signal-to-noise ratio since less common causes intrinsically have limited death counts. Consequently, cause-specific analysis must contend with much noisier data. One motivation for the MOGP approach is to explicitly enable data fusion across populations, sharing information to improve model fitting. We demonstrate significant de-noising of mortality experience that successfully captures cause-specific trends, including for causes with low data credibility. We also document the benefit of joint models to reduce model risk, that is, improved inference of model hyperparameters. As another feature, our framework can handle non-rectangular datasets, for example, countries with different period coverages. In one of our case studies, we exploit this to borrow the most recent data from other countries to update predicted domestic mortality rates.

Literature Review. One of the few works approaching cause-specific mortality modelling within a multi-population context is the recent study in Lyu *et al.* (2021). The authors introduced a nested model in the spirit of Li & Lee (2005) to jointly model major causes from three European countries by capturing the cross-cause and cross-country dependencies through common factors.

Since a given death is associated with a single cause-of-death, direct dependence among causes is not observable. As such, many researchers choose to model and forecast each cause in isolation. A variety of forecasting methods for individual causes are employed: univariate time series, such as ARIMA methods in Caselli (1996), Knudsen & McNown (1993), McNown & Rogers (1992), dynamic parametrisation in Tabeau *et al.* (1999), and least squares methods and variations of the Lee–Carter model in Caselli *et al.* (2019). To capture dependence between causes, a common approach is via copulas within the framework of dependent competing risks. The main effort is to characterise the joint distribution of survival times in terms of unobserved cause-specific mortality

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

rates; see Dimitrova *et al.* (2013), Lo & Wilke (2010) and Li & Lu (2019). Alternatively, Alai *et al.* (2018) utilised multinomial logistic regression. Outside the competing risks framework, Arnold (-Gaille) & Sherris (2013) proposed a multivariate vector error correction time series model to examine the existing cointegration relationships. Another approach is to link multiple causes through a list of clinical factors and then to apply a stochastic model to forecast these factors (Foreman *et al.*, 2018).

Several studies relied on compositional data analysis to achieve coherence in the sense that cause-specific forecasts sum to the all-cause forecast. This entails modelling the by-cause distribution of deaths, directly incorporating dependence between causes; see Bergeron-Boucher *et al.* (2017) and Kjaergaard *et al.* (2019). We refer to Wilmoth (1995), Caselli *et al.* (2019) and Tabeau *et al.* (1999) for discussions on the usefulness and limitations in using cause-specific models to project all-cause mortality.

The remainder of this paper is organised as follows. Section 2 introduces cause-of-death mortality datasets from the Human Cause-of-death Database (HCD). Section 3 describes the MOGP-ICM framework and its extensions within multinational context. Section 4 focuses on how MOGPs can maximise predictive gains over single-population models and provide insights about the projected trends of aggregate mortality. Section 5 compares the results from multi- and single-level ICM. Finally, section 6 concludes with main findings and directions for further analysis.

2. The Human Cause-of-Death Database

The HCD (HCD, 2021) provides detailed cause-specific mortality data for more than a dozen developed countries. The HCD offers three levels for the classification of causes. The short list (with 16 broad categories) and the intermediate list (103 categories) are the same across countries, while the full list is country-specific. The data for each country is organised by calendar years, age groups, gender and causes. Datasets for different countries do not line up, both due to historical availability and different timelines for updates; see Figure A.1 in Appendix A.1.

As part of their extensive and well-documented post-processing, the HCD conducted a series of bridge-coding studies to reduce disruptions in mortality trends due to changes in the International Classification of Disease (ICD). The current 10th Revision of the ICD is far more detailed with the addition of 8,000 categories compared to the 9th Revision (Anderson *et al.*, 2001). As cause-of-death records switched to the 10th Revision, death counts were shifted among some categories. To minimise such jumps, the HCD reconstructed death assignments between the old and the new ICD Revisions; this reconciliation is already part of the HCD datasets we used.

In our first case study, we analyse subcategories of cancer, available from the 103-category intermediate list. Being the leading cause of death worldwide (WHO, 2021), it is useful to understand the trends in cancer mortality for different age groups and explore the dependence between its common variations. Moreover, cancer types generally do not feature any substantial trend discontinuity due to ICD revisions (Anderson *et al.*, 2001), allowing a better assessment of model performance in terms of in-sample smoothing and out-of-sample forecasting.

For our testbed, we select five variations of cancer: lung and bronchus (LUN), colon and rectum (COL), pancreas (PAN), stomach (STM), and all other cancers (RMN) from three countries in the HCD: Czech Republic (CZE), Germany (DEU) and Poland (POL). The chosen cancer causes-of-death are common in both male and female populations, enabling us to jointly model mortality rates across cause, country and gender factors. For example, we exclude breast cancer from the study as the respective male counts are very limited. Age in the HCD is formatted as discrete 5-year age groups; we treat it as a continuous covariate encoded via the respective group average (52 for age group 50–54 years, 57 for age group 55–59 years, etc). In Figure 1, we visualise the raw log-mortality rates across different cancer types in male populations from Czech Rep., Germany and

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

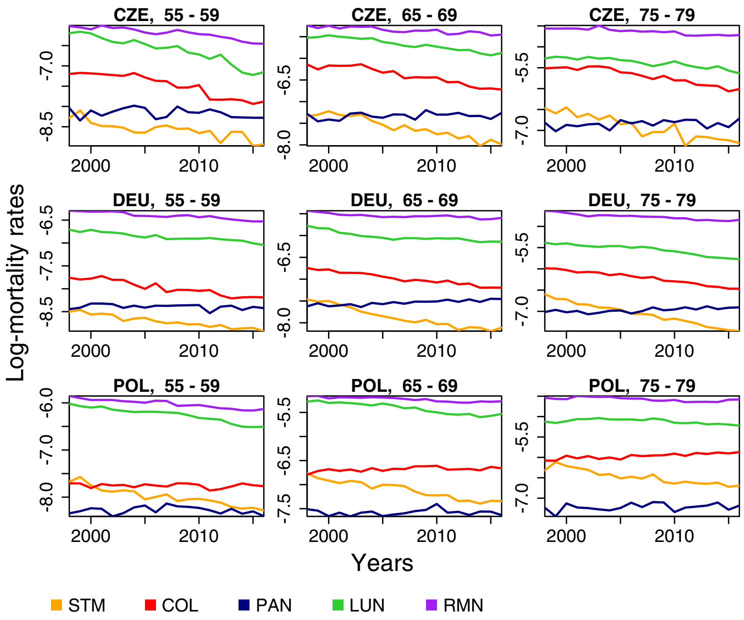


Figure 1. Log-mortality rates of selected cancers by country and age groups among males. Note the different *y*-axes in each panel. Source: HCD.

Poland. Lung and colorectal are the leading causes of cancer deaths, while pancreatic and stomach are less prevalent and exhibit more volatility. Figure A.2 in the Appendix visualises the respective patterns in age; we observe a strong convexity in age, especially for LUN.

Remark 1. Outside of the former Soviet republics, the data in HCD only goes back to the late 1990s; less than 20 years of history may be limiting for analysing the trends over time. At the same time, cause-of-death statistics are highly non-stationary due to the regular coding updates and changing medical practices. Our approach is localising in the sense that distant historical data have only a second-order impact on the outputted predictions, mitigating above concerns.

Our second case study uses a subset of the HCD data for United States, based on the CDC's National Center for Health Statistics, where we decompose all-cause mortality into its major categories. This analysis is inspired by a similar study conducted by the SOA (Boumezoued *et al.*, 2019) that employed a multivariate Lee–Carter model. The SOA study looked at 11 major causes, all age groups between 0 and 95+ years in 1999–2016 with data combined from the HCD and the Global Burden of Disease project. We restrict analysis to ages 40–69 years and years 1999–2018 and moreover to reduce the eight most common causes for the examined age groups: heart (HEA), stroke (STK), diabetes (DIA), cancers not induced by smoking (CAN), lung cancers induced by smoking (CANL), respiratory (RES), drug abuse (DRU) and all other remaining causes (RMN). The cause mapping for this case study closely follows the SOA guidance.

2.1 Stacking sub-populations

For joint modelling purposes, datasets are stacked together. Generically, we have a total of L populations, indexed by the subscript $l = 1, 2, \ldots, L$. When needed, to indicate country, cause, and gender factors, we re-index by $l = (c, s, g)$. Throughout, the two main independent variables are age and year, (x^i_{ag}, x^i_{yr}), and the observed mortality rate in the l-th population is denoted by:

$$\mu_l^i := \frac{\text{Death counts at } (x_{ag}^i, x_{yr}^i) \text{ of type } l}{\text{Exposed-to-risk counts at } (x_{ag}^i, x_{yr}^i)} = \frac{D_l^i}{E_l^i}, \tag{1}$$

where D_l^i is the number of deaths in the lth population for the ith observation with the corresponding age and year inputs. The denominator E_l^i refers to the corresponding exposed-to-risk counts. Some of the E_l^i's will be the same when we consider different causes: for a fixed country–gender combination (c, g) $E_{(c,s,g)}^i$ does not depend on s while $D_{(c,s,g)}^i$ does.

Our GP models work with log-mortality rates so that our data is

$$y_l^i = \log \mu_l^i.$$

The overall dataset is then represented as (x^i, y_l^i), $i = 1, \ldots, N_l$, $l = 1, \ldots, L$.

3. Methodology

3.1 Gaussian process regression for mortality tables

Consider the non-parametric additive regression model for the L populations to be jointly analysed:

$$y_l^i = f_l(x^i) + \epsilon_l^i, \qquad i = 1, \ldots, N, \tag{2}$$

where x^i represents an individual entry in the mortality table (indexed by age and year), y_l^i is the observed output in the lth population, and $f_l(\cdot)$, $l = 1, \ldots, L$ is the underlying latent log-mortality surface, obscured with the observation noise ϵ_l^i.

We first summarise the *spatial* structure that concerns the dependence of mortality rates as a function of age and year. Momentarily focusing on a single-output Gaussian process (SOGP) regression, we put a GP prior on the latent function $f_l \sim \mathcal{GP}(m, C)$, meaning that any finite vector $f_l(\mathbf{x}) = (f_l(x^1), \ldots, f_l(x^n))$ at n inputs follows the multivariate Gaussian distribution:

$$f_l(x^1), \ldots, f_l(x^n) \sim \mathcal{N}\big(\mathbf{m_l}(\mathbf{x}), \mathbf{C}_l(\mathbf{x},\mathbf{x})\big),$$

where $m_l(x) = \mathbb{E}[f_l(x)]$ is the mean vector of size n and $C_l(x, x') = \mathbb{E}[(f_l(x) - m_l(x))(f_l(x') - m_l(x'))]$ is the n-by-n covariance matrix. All properties of a GP are thus completely described by its mean and covariance functions. The imposition of this Gaussian structure is purely for machine learning purposes in order to leverage the extensive theory of reproducing kernel Hilbert spaces and focus the modelling efforts on the covariance kernel as a way to describe the dependence across different mortality rates.

The functions $m_l(\cdot)$ and $C_l(\cdot, \cdot)$ characterise our prior beliefs about the response surface f_l. *The covariance kernel* of the GP defines a similarity between pairs of data points. It characterises the smoothing process by determining the influence of observations on the distribution of the output. Data points that are close are expected to behave more similarly than data points that are farther away. In terms of spatial dependence on age and year, we concentrate on a common family of covariance functions known as the Matérn class, equipped with automatic relevance determination. Specifically, the Matérn-5/2 kernel defines the covariance between two mortality table entries x, x' as follows:

$$C_l(x, x') = \prod_{k \in \{ag, yr\}} \left(1 + \frac{\sqrt{5}}{\theta_{k,l}} |x_k - x_k'| + \frac{5}{3\theta_{k,l}^2} |x_k - x_k'|^2\right) \exp\left(-\frac{\sqrt{5}}{\theta_{k,l}} |x_k - x_k'|\right). \tag{3}$$

This kernel is parametrised by the age lengthscale $\theta_{ag,l}$ and the year lengthscale $\theta_{yr,l}$. Our choice of (3) is driven by the popularity of Matérn-5/2 in the GP literature, which produces twice-differentiable predictive surfaces (important for stably evaluating mortality improvement factors) but is more flexible than an infinitely differentiable kernel. See Huynh & Ludkovski (2021) for

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

further discussion of kernel choice. In short, this is a separate, non-trivial task that has some, but not drastic, effects on the results.

The mean function $m_l(x)$ describes the relevant trends in log-mortality rates. Typically in GP models, the mean is a constant; however, this is not an appropriate choice given the strong age dependence of mortality. To capture the long-term longevity features, we thus fit a parametric trend: $m_l(x) = \beta_{0,l} + \sum_{j=1}^{p} \beta_{j,l} h_j(x)$, where h_j's are given basis functions and the $\beta_{j,l}$'s are unknown coefficients. Motivated by Figure A.2 in Appendix A.2 which shows that cause-specific log-mortality tends to exhibit a concave, rather than a linear pattern in age, we consider a quadratic trend in age and linear trend in the year dimension:

$$m_l(x^i) = \beta_{0,l} + \beta_{1,l}^{ag} x_{ag}^i + \beta_{2,l}^{ag} (x_{ag}^i)^2 + \beta_{l,1}^{yr} x_{yr}^i. \tag{4}$$

We are interested in the posterior distribution for $\mathbf{f}_* \equiv f_l(\mathbf{x}_*)$ at specified inputs $\mathbf{x}_*$ given the dataset $\mathbf{y}_l = (y_l^1, \ldots, y_l^N)$, $p(\mathbf{f}_*|\mathbf{y_l})$, in other words the likelihood of the true response surface being $\mathbf{f}_*$ given what we have observed. Using $\mathrm{Cov}(y_l^i, y_l^j) = \mathrm{Cov}(f_l(x^i), f_l(x^j)) + \sigma_l^2 \delta(x^i, x^j)$ where $\delta(x^i, x^j)$ is the Kronecker delta, we have the Gaussian observation likelihood $\mathbf{y_l} \sim \mathcal{N}(\mathbf{m_l(x)}, \mathbf{C_l(x,x)} + \mathbf{\Sigma}_l)$ where the error terms are assumed to be independent and Gaussian-distributed, $(\epsilon_l^1, \ldots, \epsilon_l^N) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_l = \mathrm{diag}(\sigma_l^2))$. Incorporating the evidence from observations, as reflected in the likelihood function and the prior, we obtain the Gaussian posterior:

$$p(\mathbf{f}_*|\mathbf{y}_l) \sim \mathcal{N}(\mathbf{m}_*(\mathbf{x}_*, \mathbf{l}), \mathbf{C}_{*,\mathbf{l}}(\mathbf{x}_*, \mathbf{x}_*)).$$

The Universal Kriging equations (5)–(7) (Rasmussen & Williams, 2005, Ch 2.7) below provide not only the posterior mean $m_*(\cdot, l)$ and posterior variance $s_*^2(\cdot, l)$ but also the estimated coefficients $\boldsymbol{\beta}_l = (\beta_{1,l}, \ldots, \beta_{p,l})^T$. 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 $\mathbf{D} = (\mathbf{C} + \mathbf{\Sigma})^{-\mathbf{1}}\mathbf{H}$ where $\mathbf{C}$ is the covariance matrix $C_l(x^i, x^j)_{i,j=1}^N$. The posterior mean of $\boldsymbol{\beta}_l$ along with the predicted posterior mean $m_*(x_*, l)$ and respective variance $s_*^2(x_*, l) = C_{*,l}(x_*, x_*)$ for any input x_* are as follows:

$$\hat{\boldsymbol{\beta}}_l = (\mathbf{H}^T\mathbf{D})^{-1}\mathbf{H}^T(\mathbf{C} + \mathbf{\Sigma})^{-1}\mathbf{y}; \tag{5}$$

$$m_*(x_*, l) = \mathbf{h}(x_*)^T\hat{\boldsymbol{\beta}} + \mathbf{c}(x_*)^T(\mathbf{C} + \mathbf{\Sigma})^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}); \tag{6}$$

$$s_*^2(x_*, l) \ = (\mathbf{h}(x_*)^T - \mathbf{c}(x_*)^T\mathbf{D})^T(\mathbf{H}^T\mathbf{D})^{-1}(\mathbf{h}(x_*)^T - \mathbf{c}(x_*)^T\mathbf{D}) \tag{7}$$

where $\mathbf{c}(x_*) = (C_l(x^1, x_*), \ldots, C_l(x^N, x_*))$ is the vector of covariances between inputs in the training set and desired test input x_*. Note that the predictive distribution of observation y_l at x_* is similarly obtained as $y_l \sim \mathcal{N}(m_*(x_*, l), s_*^2(x_*, l) + \sigma_l^2)$.

Below we use $m_*(x_*, l)$ as the model prediction for the respective (log)-mortality rate of the lth population in cell x_*, and $s_*(x_*, l)$ as the corresponding posterior uncertainty which is used to obtain predictive quantiles around the former prediction.

Remark 2. The choice of a Gaussian observation likelihood in (2) is motivated by the convenience of the resulting conjugate equations that yield a Gaussian posterior. This can be generalised to a non-Gaussian (e.g. Poisson) likelihood whereby one needs to apply Laplace approximation to obtain $m_*(x_*, l)$ and $s_*^2(x_*, l)$. One can also extend to non-constant observation noise $\mathrm{Var}(\epsilon_l^i)$ to reflect varying credibility of raw observations across ages. These extensions are beyond the scope of this article.

3.2 Semi-parametric latent factor model

The vector-valued latent response variable over the age–year input space is defined as $\mathbf{f(x)} = (f_1(\mathbf{x}), \ldots, f_L(\mathbf{x}))$, where the functions $\{f_l(x)\}_{l=1}^L$ are the log-mortality surfaces for the corresponding lth population. Similar to single-population GP above, we place a GP prior over the latent function $\mathbf{f}$ such that:

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

where $\mathbf{m}$ is the mean vector function whose elements $\{m_l(x)\}_{l=1}^{L}$ are the mean functions of each output and $\mathbf{C}$ is the fused (meaning combining both spatial and cross-population terms) covariance matrix. This implies that we separate the "spatial" dependence, encoded in x from the inter-task dependence encoded in l. While x has a natural Euclidean metric that is used in (3), the population indices l are generally of factor type. Hence, to describe the dependence across L outputs, we need $L(L-1)/2$ hyperparameters $C_{lk}^{(f)}$, $1 \le l, k \le L$ which becomes inefficient and unstable beyond 3-4 populations. Thus, with an eye towards reducing the number of hyperparameters, additional structure is needed for $C^{(f)}$.

In the SLFM dating back to Teh *et al.* (2005), each output $f_l(x)$ is assumed to be a linear combination of Q latent functions:

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

where $u_q(x)$'s are independent realisations from GP priors with distinct covariances $C_q^{(u)}(x, x')$ and $a_{l,q} \in \mathbb{R}$'s are the factor loadings $(q = 1, \ldots, Q)$, considered part of the kernel hyperparameter space Θ. Thus, the semiparametric name of the model comes from the combination of a nonparametric component (several GPs) and a parametric linear mixing of the functions $u_q(x)$.

The role of $Q \le L$ is to achieve dimension reduction for the correlation structure across the f_l's. Let $\mathbf{a}_q = (a_{1,q}, \ldots, a_{L,q})^T$ be the vector representing the collection of coefficients associated with the qth latent function across the L outputs. Then, the covariance of the vector-valued function $\mathbf{f}(\mathbf{x}) = \sum_{q=1}^{Q} \mathbf{a}_q u_q(\mathbf{x})$ is as follows:

$$\begin{aligned} \mathrm{Cov}(\mathbf{f}(x), \mathbf{f}(x')) &= \sum_{q=1}^{Q} (\mathbf{a}_q \mathbf{a}_q^T) \otimes C_q^{(u)}(x, x') \\ &= \sum_{q=1}^{Q} (A_q A_q^T) \otimes C_q^{(u)}(x, x') = \sum_{q=1}^{Q} B_q \otimes C_q^{(u)}(x, x') \end{aligned} \tag{9}$$

where $\otimes$ symbolises the Kronecker product, $A_q = \mathbf{a}_q = (a_{1,q}, \ldots, a_{L,q})^T$ and each B_q has rank one.

3.3 Intrinsic coregionalisation model

Similar to SLFM, ICM assumes each output function $f_l(x)$ is generated from a common pool of Q latent functions, cf. (8). However, the latent $u_q(x)$ all share the *same* GP prior with the covariance kernel $C^{(u)}(x, x')$. Then, the covariance for $\mathbf{f}(x)$ is

$$\mathrm{Cov}(\mathbf{f}(x), \mathbf{f}(x')) = \left(\sum_{q=1}^{Q} \mathbf{a}_q \mathbf{a}_q^T \right) \otimes \mathrm{Cov}(u_q(x), u_q(x')) = B \otimes C^{(u)}(x, x'), \tag{10}$$

where $B = AA^T \in \mathbb{R}^{L \times L}$ has rank Q. In other words, the cross-output correlation is of rank $Q \le L$, while the spatial covariance of each output is the same. The lth element in the diagonal of the cross-covariance matrix B (or $\sum_{q=1}^{Q} B_q$ in SLFM) represents the process variance of $f_l(\cdot)$. Since $B = \sum_{q=1}^{Q} \mathbf{a}_q \mathbf{a}_q^T$, the individual entries are $B_{l,k} = \sum_{q=1}^{Q} a_{l,q} a_{k,q}$ and the diagonals are $B_{l,l} = \eta_l^2 = \sum_{q=1}^{Q} a_{l,q}^2$, $1 \le l \le L$. We can similarly infer the correlation matrix $R = (r_{l_1,l_2})$ between population

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

l_1 and l_2 $(1 \leq l_1, l_2 \leq L)$; for ICM, it is

$$r_{l_1,l_2} := \frac{\mathrm{Cov}(f_{l_1}(x), f_{l_2}(x))}{\sqrt{\mathrm{Var}(f_{l_1}(x)) \times \mathrm{Var}(f_{l_2}(x))}} = \frac{B_{l_1,l_2}}{\sqrt{B_{l_1,l_1} B_{l_2,l_2}}} = \frac{\sum_{q=1}^{Q} a_{l_1,q} a_{l_2,q}}{\sqrt{\left(\sum_{q=1}^{Q} a_{l_1,q}^2\right)\left(\sum_{q=1}^{Q} a_{l_2,q}^2\right)}}. \tag{11}$$

In both SLFM and ICM, the fused covariance kernel belongs to the class of separable kernels (Alvarez *et al.* 2012) and decouples using the Kronecker product into: (1) the coregionalisation matrices B_q that measure the interaction between different outputs and (2) the spatial covariance over age–year dimensions $C_q^{(u)}(x, x')$, cf. Equations (9) & (10). In ICM, all L populations share the same spatial covariance kernel $C^{(u)}(x, x')$. Such assumption of a common spatial covariance over age–year inputs links to the concept of commonality in the mortality structure (but not levels) across populations. Compared to ICM, SLFM offers more flexibility at the cost of adding more hyperparameters: the total number of kernel hyperparameters in the fused covariance matrix **C** of SLFM is $QL + 2Q$ vis-a-vis $QL + 2$ hyperparameters in the ICM.

Selecting rank *Q*. As Q is not one of the hyperparameters to be optimised, ad hoc ways are needed to pick it. We use the Bayesian information criterion (BIC) to select rank Q that produces the most parsimonious model; see Williams *et al.* (2009) and Huynh & Ludkovski (2021). As discussed in Bonilla *et al.* (2008), taking $Q < L$ in ICM corresponds to finding a rank-Q approximation (based on an incomplete Cholesky decomposition) to the full-rank $C^{(f)}$. A similar interpretation holds for SLFM and the respective B_q's. While attractive for dimension reduction and computational speed up, low Q may not be adequate to describe the overall dependence structure and hence clashes with the original goal of capturing the variability present in the fused mortality dataset. In particular, we observe that BIC tends to select $Q \in \{2, 3\}$ which may be too small for $L \geq 5$. Based on our case studies, we recommend $Q \in \{3, 4, 5\}$ for maximising predictive performance.

3.4 Multi-level ICM for scalable GPs

In the situation when we have multi-dimensional factor inputs (e.g. cause and gender together), one approach is to combine all factor inputs into a single covariate with L distinct outputs prior to applying ICM or SLFM. When L grows large, ICM becomes less feasible due to its time complexity $\mathcal{O}(N^3 L Q^2)$ (Bonilla *et al.*, 2008). In this section, we develop the structured Kronecker product kernel (multi-level ICM in Liu *et al.*, 2020; Zhe *et al.*, 2019) to mitigate this scalability issue in GP. The structured covariance kernel exploits the fact that mortality tables are gridded along each factor dimension.

We express the total number of outputs L as the product across P types of categorical inputs, $L = \prod_{p=1}^{P} L_p$, where L_p is the number of levels within the pth categorical input. We then decompose the cross-population covariance $\tilde{B}$ as the Kronecker product:

$$\tilde{B} = \bigotimes_{p=1}^{P} \tilde{B}_p \tag{12}$$

where $\tilde{B}_p$, $1 \leq p \leq P$ refers to the cross-covariance matrix between sub-populations within the pth categorical input, taken to have rank $Q_p \leq L_p$. Directly marginalising the cross-covariance matrices yields a convenient interpretation of the correlation between sub-populations within a factor input and moreover allows for separate estimation of each cross-covariance sub-matrix $\tilde{B}_p$, $1 \leq p \leq P$.

The multi-level ICM set-up implies that each output $f_l(x)$ is the weighted combination of $Q_1 \times \ldots \times Q_P$ independent latent functions, all with the spatial covariance kernel $C^{(u)}(x, x')$:

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

$$f_l(x) = \sum_{j=1}^{Q_1 \times \ldots \times Q_P} a_{l,j} u_j(x). \tag{13}$$

The improvement in scalability of the multi-level ICM can be analysed via the ranks Q_p of the cross-covariance sub-matrices. Thanks to the Kronecker product's property, $\text{rank}(\tilde{B}) = \prod_{p=1}^{P} \text{rank}(\tilde{B}_p) = \prod_{p=1}^{P} Q_p$ and using Cholesky decomposition, Equation (12) can be rewritten as:

$$\tilde{B} = \tilde{A}\tilde{A}^T = \bigotimes_{p=1}^{P} \tilde{B}_p = \bigotimes_{p=1}^{P} (\tilde{A}_p \tilde{A}_p^T) = \left(\bigotimes_{p=1}^{P} \tilde{A}_p \right) \left(\bigotimes_{p=1}^{P} \tilde{A}_p \right)^T, \tag{14}$$

where $\tilde{A}_p = (\mathbf{a}_1^p, \ldots, \mathbf{a}_{Q_p}^p)$, $1 \le p \le P$, and each vector $\mathbf{a}_k^p = (a_{1,k}^p, \ldots, a_{L_p,k}^p)^T$, $(1 \le k \le Q_p)$ represents the collection of scalar coefficients associated with the kth latent function across L_p sub-populations in the pth categorical input. Thus, the number of hyperparameters required to estimate the cross-covariance $\tilde{B}$ is $\sum_{p=1}^{P} Q_p L_p$, which can be much lower than for single-level ICM when L is large. Note that when $\text{rank}(B) < \text{rank}(\tilde{B})$, the multi-level ICM utilises more latent functions to generate the model outputs, compensating for the imposed structure in $\tilde{B}$ in (14). In terms of overall complexity, multi-level ICM requires $\mathcal{O}(N^3(\sum_p L_p Q_p^2))$ time compared to $\mathcal{O}(N^3 L Q^2)$ for single-level ICM.

Remark 3. In the typical situation, L_p's are small and so it is feasible to consider full-rank multi-level ICM, that is, $Q_p = L_p$. Otherwise, (14) allows to exploit simultaneously the Kronecker product structure, as well as the low-rank approximation. See Table 2 for results on the impact of Q_p values in multi-level ICM.

3.5 MOGP hyperparameters

To implement a GP model requires specifying its hyperparameters. Note that actual inference reduces to linear algebraic formulas in (6)–(7), and the modeling task is to learn the spatial covariance, namely the mean and kernel functions.

Mean function: We make the prior $m_l(x)$ to be population-specific in order to maximise model flexibility in describing the mortality trend of each population. Thus, we have $3L+1$ coefficients $\boldsymbol{\beta} = (\beta_0, \beta_{1,l}^{ag}, \beta_{2,l}^{ag}, \beta_{l,1}^{yr} : l = 1, \ldots, L)$, cf. (4).

Observation Likelihood: We assume a constant observation noise within each population $\sigma_l = \text{StDev}(\epsilon_l^i)$. This accounts for heterogeneous characteristics when observations from multiple populations are combined; in particular, σ_l is smaller for larger populations and for more prevalent causes (Huynh *et al.*, 2020). The σ_l's are estimated via maximum likelihood along with all other hyperparameters. More advanced GP models that either employ Poisson likelihood or infer input-dependent non-parametric $\sigma_l(x^i)$ are possible but require additional coding and are beyond the scope of this work.

Estimating Hyperparameters: For (multi-level) ICM, the set of hyperparameters is $\Theta = (\theta_{ag}, \theta_{yr}, (a_{l,q}), (\sigma_l^2), \boldsymbol{\beta})$; for SLFM, it is $\Theta = \big((\theta_{ag,q}), (\theta_{yr,q}), (a_{l,q}), (\sigma_l^2), \boldsymbol{\beta}\big)$. We use the R package kergp (Deville *et al.*, 2019) to carry out the respective maximum likelihood estimation via Kronecker decompositions. Alternatively, to account for model risk and offer more robust results, one could employ a fully Bayesian hyperparameter inference. This could be done with the Stan software (Carpenter *et al.*, 2017) but is beyond the scope of this work.

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

3.6 Performance metrics

Given a test set of observed $y_*(x_*, l)$'s, we evaluate the effectiveness of different models using two metrics. First, we employ the mean absolute percentage error (APE) to examine the discrepancy between the observed and predicted outputs:

$$\text{APE}(y_*, m_*) := \left| \frac{y_* - m_*(x_*)}{y_*} \right| \tag{15}$$

where $y_*(x_*, l)$ is the observed value at test input x_* in the lth population and $m_*(x_*, l)$ is the predicted log-mortality rate. Note that APE is scale-independent, enabling us to compare model performance across populations with different exposures.

We also use the continuous ranked probability score (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 $y_*(x_*, l)$ which allows a more detailed uncertainty quantification beyond just looking at the predictive mean $m_*(x_*, l)$. CRPS assesses the closeness of the realised outcome $y_*(x_*, l)$ relative to its predictive distribution $F_*(\,\cdot\,; x_*)$ which in the GP contest is Gaussian and leads to

$$\begin{aligned} \text{CRPS}(F_*, y_*) &:= \int_{\mathbb{R}} \left[F_*(z) - \mathbb{1}_{\{z \geq y_*\}} \right]^2 \, dz \\ &= \sqrt{s_*^2(x_*) + \sigma_l^2}\left[\tilde{y}_*(2\Phi(\tilde{y}_*) - 1) + 2\phi(\tilde{y}_*) - \frac{1}{\sqrt{\pi}}\right], \quad \tilde{y}_* := \frac{y_* - m_*(x_*)}{s_*^2(x_*) + \sigma_l^2} \end{aligned} \tag{16}$$

where $\phi(\cdot)$, $\Phi(\cdot)$ are the standard Gaussian density and cdf. Observe how CRPS penalises both bias $(2\Phi(\tilde{y}^*) - 1)$ and excessive predictive variance.

We average both APE and CRPS across a test set of age–year–population inputs. The resulting mean APE is interpreted as a normalised relative predictive error, and mean CRPS as the squared difference between the forecasted and the empirical cumulative distribution functions. Models with lower mean APE/CRPS are judged to have a better fit.

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. The raw annual percentage mortality improvement is $MI_l^{obs}(x) := 1 - \frac{\exp\left(y_l(x_{ag};x_{yr})\right)}{\exp\left(y_l(x_{ag};x_{yr}-1)\right)}$. The smoothed improvement factor is obtained by substituting in the GP posterior means m_*'s:

$$MI_l^{GP}(x) := \left[1 - \frac{\exp\left(m_*(x_{ag}; x_{yr}, l)\right)}{\exp\left(m_*(x_{ag}; x_{yr} - 1, l)\right)} \right]. \tag{17}$$

4. Modelling Multiple Causes of Death

To understand the behaviour of age–year-specific mortality across different causes of death, we begin by generating MOGP models for cancer variants. Using the HCD database, we fit both ICM and SLFM and assess their performance in three European countries (thus, populations are indexed by cause only, $l \equiv s$, and we build independent models for males in each country). We test the resulting predictive performance by computing APE and CRPS for 1-year-ahead mortality forecasts in three separate test sets, using SOGP as the baseline. All models, including SOGP models, are trained on the same ages from 50 to 84 years (seven age groups) and three overlapping periods: 1998–2013 for the 2014 prediction, 1999–2014 for 2015 prediction and lastly 2000–2015 for 2016 prediction. We report APE and CRPS for all-cancer log-mortality rates, since some of the cancer variations such as stomach and pancreatic have relatively few (and therefore more noisy) recorded deaths. The differences in APE and CRPS between MOGP and SOGP models are

Table 1. Comparison between MOGP ICM/SLFM with different ranks Q, reported as the relative improvement in APE and CRPS of MOGP vis-a-vis SOGP. The reported values are medians of 1-year-out aggregated all-cancer forecasts for age groups 50–84 years based on three training periods: 1998–2013 (predict 2014), 1999–2014 (predict 2015) and 2000–2015 (predict 2016). Both ICM and SLFM are fitted on five cancer types, male populations in each selected country.

		# Kernel	Czech Rep		Germany		Poland	
Cause ($L=5$)		Hyperparameters	APE	CRPS	APE	CRPS	APE	CRPS
ICM	$Q=2$	12	19.16	30.18	3.86	15.05	12.85	17.72
	$Q=3$	17	17.69	19.08	6.53	15.16	13.46	18.64
	$Q=4$	22	22.31	21.43	6.30	14.58	17.02	11.01
SLFM	$Q=2$	14	18.04	29.49	5.09	13.51	10.82	16.05
	$Q=3$	21	22.87	23.72	2.78	13.55	4.11	10.65
	$Q=4$	28	16.97	18.24	8.75	16.73	5.13	3.81

expressed as the 3-year median percentage improvement over SOGP models. Positive improvement means joint models have smaller mean APE/CRPS values. To compute all-cancer CRPS, we leverage the closed-form expression of the MOGP multivariate predictive distribution in (6)–(7) to simulate the forecast distribution of all-cancer log-mortality for each age group in the data. To do so, we first draw (5×10^5) stochastic samples of the joint $\mathbf{f}_*(x_*)$ across all the cancer types. After exponentiating and summing, we then obtain corresponding samples of (unlogged) all-cancer mortality rates.

Table 1 shows the 3-year median improvement in MAPE and CRPS for multi-cause ICM and SLFM over SOGP models. Overall, joint models produce more accurate mean forecasts (positive improvement in APE) with higher credibility (positive improvement in CRPS). We observe the opportunity to borrow information across different cancers to better estimate the kernel hyperparameters. As a result, joint models can describe important trends for individual cancers, leading to the reduction in disparity between the predicted values and the observed all-cancer mortality in the test sets.

4.1 Commonalities in cause-specific mortality surfaces

The main difference between ICM and SLFM is the underlying assumption about the latent factors $u_q(\cdot)$. ICM assumes that all factors have the same lengthscales θ_{ag}, θ_{yr} and is therefore appropriate for modelling homogenous mortality surfaces. SLFM is more general and fits distinct $\theta_{ag,q}, \theta_{yr,q}$; it is expected to perform better when the different surfaces exhibit heterogeneity (e.g. different degree of correlation across age). We examine this commonality assumption in Figure 2 by displaying side by side the mortality improvement factors of the various cancers derived from multi-cause ICM and SLFM. The BIC selection criterion yields $Q=2$ for both ICM and SLFM. In this case study, the results from SLFM are almost identical to ICM predictions. Therefore, the assumption of sharing the spatial kernel over age–year inputs across the considered cancer types is plausible. This conclusion is reinforced by Table A.1 in Appendix A which shows that the inferred lengthscales $\theta_{ag,q}, \theta_{yr,q}$ for SLFM are very similar for $q = 1, 2$ across all three countries. In other words, both of the latent factors learned by SLFM have similar age–year spatial dependence, and so there is little loss of fidelity in a priori forcing them to be equal, as is done in ICM. Indeed, the lengthscales in SLFM are close to the ICM ones.

We can also inspect Figure 2 for insights about the relative mortality improvement trends of different cancers. Stomach cancer has the largest improvement rates for most age groups in all three countries. Decline in stomach cancer incidence tends to be associated with economic improvements resulting in healthier diet, better food preservation, clean water supply, etc. We also observe the increasing improvement trend of lung cancer among age groups below 60 years in Czech Rep. and Poland, reflecting lower smoking rates in birth cohorts after WWII. Czech

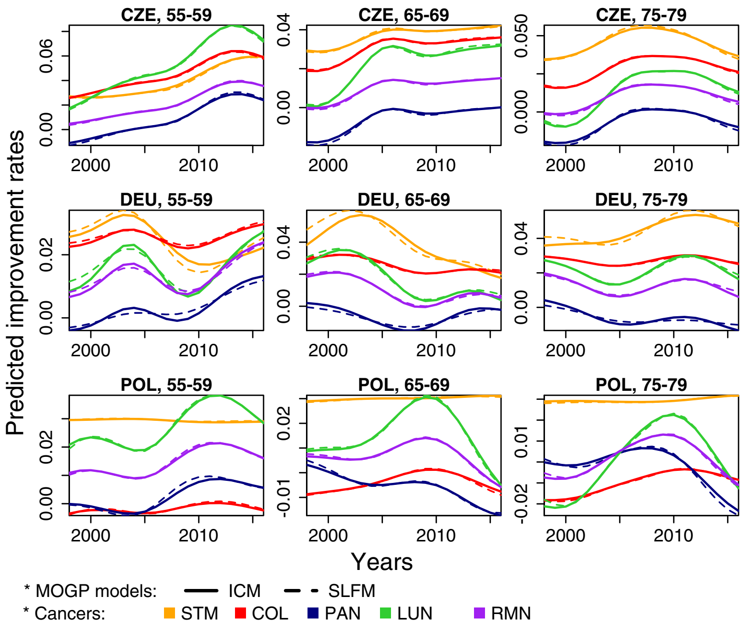
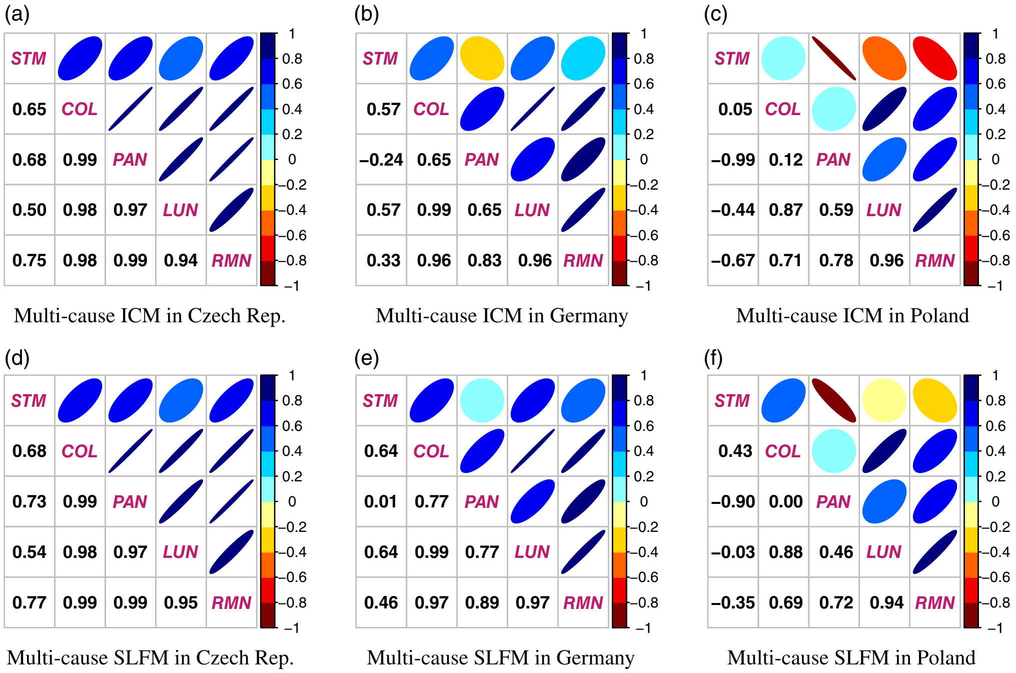


Figure 2. The predicted YoY improvement rates derived from multi-cause GP models by country and age groups among male observations. In each country, MOGP-ICM ($Q = 2$) and SLFM ($Q = 2$) are fitted on ages 50–84 years (seven groups), years 1998–2016 over five cancer variations: stomach, colorectal, pancreatic, lung and remaining types.

males experienced large increase in the improvement rates in most cancers. In Germany, the improvement rates have been rising among age groups below 60 years but slowing down among older age groups. In Polish males, except stomach cancer, the improvement trends increased until early 2010s and then significantly declined, displaying the impact of an ageing population and an increase in lifestyle exposure to risk factors for cancers (Wojtys *et al.*, 2014). The incidence of stomach cancer has been flat over time, consistent with the finding in Arash *et al.* (2020).

Figure 3. visualises the inferred cross-cause correlation matrices $(r_{l_1,l_2}{:}1 \le l_1, l_2 \le 5)$. Both ICM and SLFM document a global positive association among mortality rates from these cancers in each country. It is consistent with strong resemblance in the mortality improvement trends between cancers in Figure 2. Since cancers within a given population share common risk factors, innovations in early detection and advancements in treatment for one cancer are likely to have positive effects on other cancer variations. Negative correlation $r_{l_1,l_2} < 0$ reflects opposite trends, for example, stomach and pancreas cancers in Poland; this may be due to a competing risks context.

For a different take on cause-of-death commonality, we applied the multi-cause GP ICM model to jointly model $L = 8$ top causes in the HCD US dataset, separately for each gender. Using BIC as the criterion, models with rank $Q = 6$ (for eight populations) yield the largest BICs for both males and females. This indicates a higher degree of heterogeneity in these larger cause groupings, compared to $Q = 2$ across five cancer causes in the previous study. Thus, the models employ more latent functions $u_q(\cdot)$ to adequately capture the total variability in the joint datasets. The inferred cross-cause correlation matrices are displayed in Figure B.2 in Appendix B. Overall, our results are in agreement with the SOA study (Boumezoued *et al.*, 2019), for example, confirming that there are moderate positive associations between most causes. The strongest correlations are found between heart disease and stroke, and between heart and drug overdose. As might be expected, there is little correlation between remaining causes (RMN) and most other categories.



Multi-cause ICM in Czech Rep.

Multi-cause ICM in Germany

Multi-cause ICM in Poland

Multi-cause SLFM in Czech Rep.

Multi-cause SLFM in Germany

Multi-cause SLFM in Poland

Figure 3. Correlation matrices R derived from multi-cause GP models; rank $Q = 2$ is chosen for both ICM and SLFM. In each country, the MOGP model is fitted on males aged 50–84 years (seven groups) during years 1998–2016, over five cancer variations: stomach, colorectal, pancreatic, lung and remaining types.

Some of the correlations vary between genders, possibly due to strong observation noise in less common causes.

A complementary way to examine commonalities across causes is to inspect the inferred factor loadings $a_{l,q}$. Populations that have similar loadings will be highly correlated. Figure A.3 in Appendix A displays the factor loadings in a country–cause SLFM with both $Q = 2$ and $Q = 3$ latent factors u_q. We observe that primary clustering is by causes rather than by countries. Some outliers, such as STM in Czech R., can also be seen and suggest idiosyncratic behaviour of the respective mortality surface. Less separation (and limited interpretation of $a_{l,i}$) is observed when using only $Q = 2$ latent factors, which indirectly suggests that $Q = 3$ is preferable.

4.2 Aggregating by-cause models

An important motivation for our work was to use by-cause analysis to make more precise conclusions about all-cause mortality. For example, the mortality trends of individual cancers give insights into the respective all-cancer mortality trends. Figure 4 shows the results from a multi-level country–cause GP ICM. It visualises the aggregated predictive distribution of all-cancer log-mortality observations, $y_l(x_*)$, for male populations by country and age group, using shading to denote predictive quantiles. Note that since the model did not use all-cancer mortality during training, the fact that the predictive in-sample bands closely match the historical movement of all-cancer mortality data is a validation of a successful by-cause analysis.

The effort in fighting cancer has transpired in all three countries, but the improvement is not uniform. Despite Czech Republic and Poland being socio-economically similar, Czech Rep. has a faster improvement pace than Poland. Although Germany continues to have the lowest log-mortality rates across all age groups, the Czech Rep. has drastically closed this gap in the last decade; see especially the left panel of Figure 4 (age group 55–59 years). The main driver

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

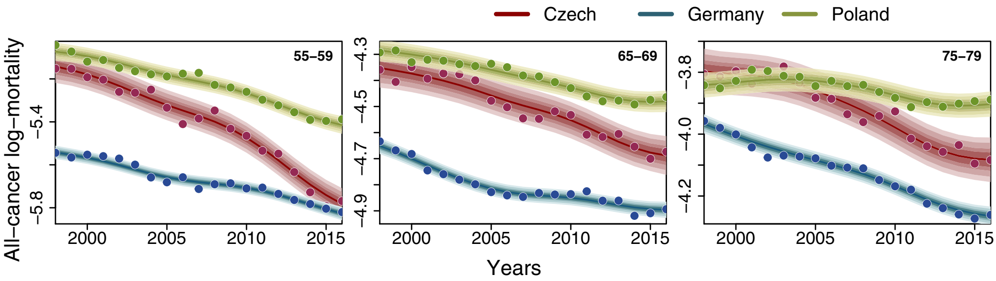


Figure 4. Predictive distribution of all-cancer log-mortality rates for different age groups in three countries via multi-level country–cause ICM. The joint model is fitted on males, age groups between 50 and 84 years, years 1998–2016 across three countries ($Q_{ctry} = 3$) and five cancers ($Q_{caus} = 5$). Shading indicates the 60%, 80%, 95% and 99% predictive quantile bands.

is the rapid improvements in all common cancers in Czech Rep., for example, the mortality improvement factor for LUN being recently more than double compared to Germany, cf. Figure 2.

4.3 Trends in US top-level causes

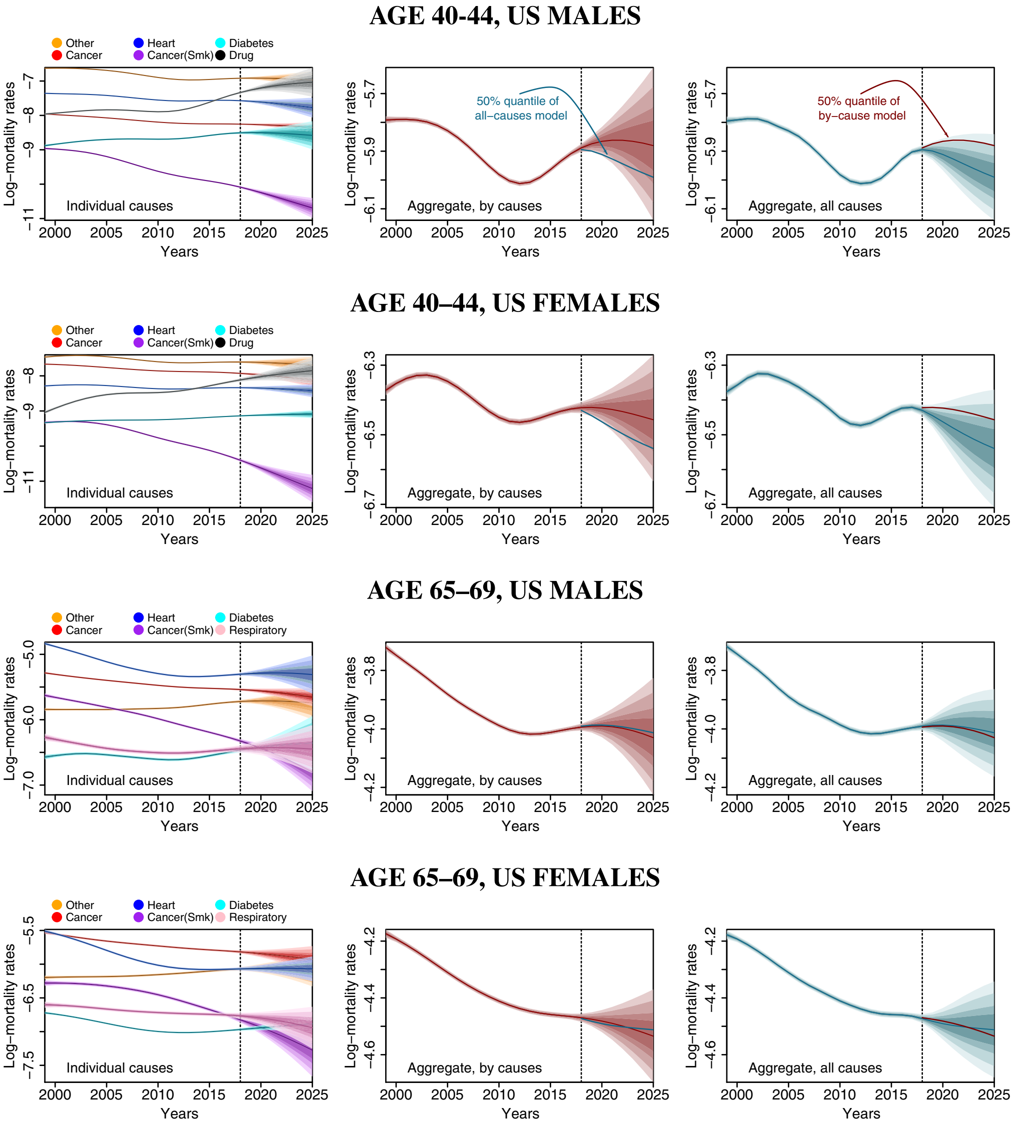
Figure 5 presents the in-sample posterior distribution of log-mortality rates $f_l(x_*)$ during 1999–2018 along with the projected trends up to 2025 for two age groups, in both male and female US populations. The left panels display the predictive trends for the top six causes in each age group. We observe improvements in most causes and age groups; the largest improvement being in CANL. In contrast, drug abuse deaths are rising rapidly among young age groups (e.g. age 40–44 years); see the two upper-left panels of Figure 5. In the older age groups, mortality from heart disease experienced large declines in early 2000s but essentially flattened out after 2015. Boumezoued *et al.* (2019) emphasised the need for break-point detection to improve forecasts of such causes whose trends change over time. In the MOGP framework, the forecasts are driven by the most recent data and get automatically gradually adjusted if trends shift. So, for example, we do not need to do anything further to achieve the slow pace of future HEA improvement as shown in Figure 5.

The middle and right columns in Figure 5 compare the aggregate all-cause projections for the US population. We witness the pessimism of the aggregate projected trends based on the by-cause models compared to an all-cause SOGP (right column), especially in the younger ages. This discrepancy, driven by the growing importance of causes with increasing trends, such as drug overdose, highlights the additional insights from by-cause modelling. The pessimism of by-cause analysis was first mentioned in Wilmoth (1995) and re-iterated in Boumezoued *et al.* (2019). For older age groups, the underlying dynamics among common causes are more stable and all-cause and by-cause forecasts are broadly similar.

Note that compared to Figure 4, the GP posterior uncertainty bands widen dramatically as we go from in-sample (up to year 2018) to extrapolating for years 2019–2025. This reflects the data-driven nature of GP forecasts which intrinsically leads to low uncertainty for in-sample smoothing and widening uncertainty as predictions are made further into the future. This phenomenon gets amplified as we add up the by-cause forecasts to obtain all-cause predictions and witness the wider band in the middle panels of Figure 5 relative to the right panel based on a SOGP.

As discussed in Huynh & Ludkovski (2021), MOGPs are well suited to generate expert-based projections. This is a useful feature to have given that by default, projections are driven by the historical trends that might not continue in the future. For instance, the increasing trend of drug abuse is largely fuelled by the opioid epidemic. Assuming this crisis is addressed in the future, and the projected DRU mortality should be adjusted downwards. In MOGP, this can be achieved by modifying the year trend in $m(\cdot)$. Figure B.1 in the Appendix displays an illustration where the

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press



AGE 40-44, US MALES

AGE 40–44, US FEMALES

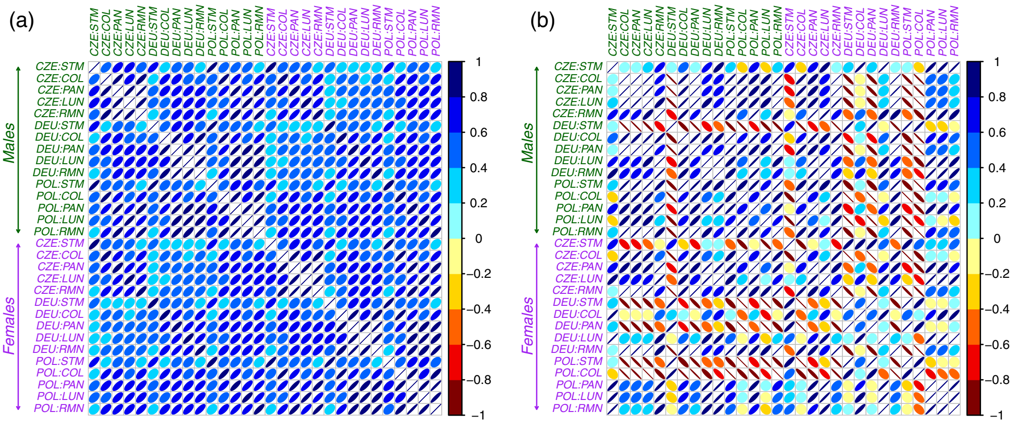
AGE 65–69, US MALES

AGE 65–69, US FEMALES

Figure 5. Posterior distribution of true log-mortality rates for US males and females for the 40–44 and 65–69 years of age groups. In each row, left panel shows the posterior quantiles for top six individual causes (via multi-cause GP), middle panel shows aggregate log-mortality trends via by-cause model (multi-cause GP) and right panel shows aggregate log-mortality trends via all-cause model (single-output GP). All models are fitted on ages 40–69 years and years 1999–2018. The vertical lines indicate the boundary between in-sample (1999–2018) and out-of-sample forecast (2019–2025). Shading indicates the 60%, 80%, 95% and 99% predictive quantile bands. We further overlay the 50% quantile (predictive median) of by-cause and all-cause models for the out-of-sample period for convenient comparison.

trend of drug abuse is reduced by one-third (through lowering the year effect β^{yr} by one-third) of the original pace for both the male and female populations. The resulting adjusted forecast for aggregate mortality gets closer to that from the all-cause model, reducing the level of pessimism we have observed earlier among the US young population. Another potential adjustment could be for smoking-induced cancer, where the MOGP models extrapolate the historical trend of rapid longevity gains. However, the SOA report (Boumezoued *et al.*, 2019) suggests that this pace might not take place in the intermediate term, lowering aggregate mortality gains.

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press



Multi-level ICM ($Q_{ctry} = 3,\ Q_{caus} = 5,\ Q_{gndr} = 2$) with 40 hyperparameters

Single-level ICM ($Q = 2$) with 62 hyperparameters

Figure 6. Cross-correlation matrices of MOGP models that incorporate country, cause and gender as categorical inputs. Thirty total populations.

5. Cause-of-Death Joint Modelling in a Multinational Context

We proceed to consider simultaneously 30 populations in the cancer variations case study, arranged by the three factor inputs of cause ($L_{caus} = 5$), country ($L_{ctry} = 3$) and gender ($L_{gndr} = 2$).

5.1 Multi-level versus single-level correlation structure

Figure 6 displays the inferred correlation structure $r_{l_1,l_2} : 1 \le l_1, l_2 \le 30$ across the above 30 populations. On the left, we fit a multi-level ICM ($Q_{ctry} = 3,\ Q_{caus} = 5,\ Q_{gndr} = 2$) and on the right a single-level ICM ($Q = 2$). Both models are fitted on age groups 50–84 years, years 1998–2016 for three countries, five cancers and two genders. Observe that the multi-level model has $\sum_p Q_p L_p + 2 = 3 \times 3 + 5 \times 5 + 2 \times 2 + 2 = 38$ hyperparameters compared to $QL + 2 = 62$ in the single-level model. The right panel does not display any recognisable structure in the inferred correlations because the model is not aware of the different factor dimensions; after dimension reduction, the marginal associations between sub-populations within the original factor inputs are no longer accessible. In contrast, the multi-level model enforces a block structure in the correlation matrix R; see Figure 6(a). Recall that the correlation sub-matrices for each factor input are estimated separately and the Kronecker product structure implies that we can read off the correlation among any combination of factors. Figure 7 displays the derived sub cross-correlation matrices from the above three-level country–cause–gender GP model. One can then multiply these factor-based correlations to get the total correlation $r_{(c_1,s_1,g_1),(c_2,s_2,g_2)} = \prod_{p \in \{c,s,g\}} r_{p_1,p_2}$ in Figure 6(a). For example, the correlation of cancer mortality between Czech-lung-male and Polish-pancreas-male is $r_{CZE,\,POL} \times r_{LUN,\,PAN} \times r_{MAL,\,MAL} = 0.88 \times 0.91 \times 1 \approx 0.80$. This is in fact also the correlation between Czech-lung-female and Polish-pancreas-female ($r_{CZE,\,POL} \times r_{LUN,\,PAN} \times r_{FEM,\,FEM}$) mortality. The limited number of deaths in some causes explains differences in the correlation matrices between single- and multi-level ICM. For example, the correlations between Polish-stomach-female and other cancer types are mostly negatively correlated in ICM but (mildly) positively correlated in the multi-level ICM.

5.2 Model selection

Next, we compare the predictive performance of multi-level versus single-level MOGP-ICM. Following the same set-up as Table 1, Table 2 shows the 3-year median improvement in APE

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

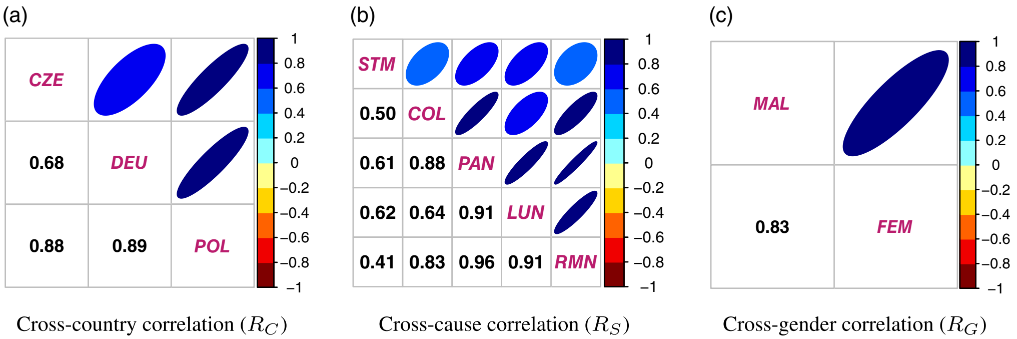


Figure 7. Cross-correlation matrices derived from a three-level country–cause–gender multi-level ICM model ($Q_{ctry} = 3$, $Q_{caus} = 5$, $Q_{gndr} = 2$), fitted on age groups 50–84 years and years 1998–2016.

and CRPS for different joint models relative to SOGP models in a multinational context. Joint models produce more accurate mean forecasts with higher credibility, but the predictive gains are not uniform across countries. The largest improvements from multi-level ICM are in Czech Republic; Czech raw data have lower credibility due to its smaller exposures; thus, there is more opportunity for data fusion. The results further validate our approach of modelling cause-specific mortality across populations: models that incorporate information from foreign countries (e.g. country–cause GP) have larger predictive improvements compared to Table 1. In Poland due to the recent structural break encountered in mid-2010s (see Figure 4), the performance of aggregated country–cause MOGP is consistently worse than that of all-cause SOGP; this issue is rectified for country–cause–gender MOGPs. We expect this issue to self-correct in a couple of years if the new trend persists.

In contrast to section 4 where the performance of ICM and SLFM was almost identical, with multiple input factors ICM usually outperforms SLFM. Given the strong commonality in mortality trends across cancer types, a shared age–year covariance kernel appears preferable for information fusion. The comparison between single- and multi-level ICM depends on the number of populations to model. In country–cause–gender setting with $L = 30$ populations, the multi-level ICM ($Q_{ctry} = 3$, $Q_{caus} = 5$, $Q_{gndr} = 2$) yields better mean APE and CRPS than single-level ICM and moreover uses fewer hyperparameters. For country–cause setting, the performance is comparable. Note that to make an apples-to-apples comparison between single- and multi-level ICM, one ought to equalise their number of hyperparameters, rather than the Q's.

Table 2 also shows the impact of the latent ranks Q and Q_p's. For country–cause, $Q = 4$ tends to yield the best results in single-level models, but for country–cause–gender, ICM with $Q = 3$ performs consistently worse than $Q = 2$, presumably due to unstable estimates of the more than 90 underlying hyperparameters. For multi-level ICM, we generally find that full-rank $Q_p = L_p$ works best, although low-rank set-ups $Q_p < L_p$ also yield good predictive performance, indicating the opportunity to shrink even further the number of kernel hyperparameters.

Remark 4. It can be seen in Table 2 that joint models do not necessarily provide more accurate forecasts. Successful data fusion requires combining similar datasets. In our example above, Germany, though neighbouring Poland and Czech Rep., has rather different demographics and so is arguably not a good candidate for data fusion. The chosen case study is driven by data availability in the HCD and is intended to be illustrative, rather than prescriptive. The broader task of selecting what countries to group together and how to obtain the most accurate predictive distribution is beyond our scope. The same remark applies when comparing country–cause–gender vis-a-vis country–cause models: it is not a priori clear whether joint modelling of both genders is beneficial, and the presented results are mixed on that issue. Including gender doubles the number of populations which requires more hyperparameters and might apparently degrade performance

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

Table 2. Comparison between MOGP models with different ranks Q in terms of APE and CRPS metrics. The reported values are median relative improvements of MOGP versus SOGP of 1-year-out aggregated all-cancer forecasts for age groups 50–84 years based on three training periods: 1998–2013 (predict 2014), 1999–2014 (predict 2015) and 2000–2015 (predict 2016). Top half: MOGP models fitted on five cancer types and three countries. Bottom half: MOGP models fitted on five cancer types, three countries and two genders.

Country+Cause ($L=15$)		# Kernel	Czech Rep.		Germany		Poland	
		Hyperparameters	APE	CRPS	APE	CRPS	APE	CRPS
ICM	$Q=2$	32	35.54	30.74	12.02	21.06	−40.44	−59.47
	$Q=3$	47	45.51	50.07	8.50	20.35	−37.09	−38.48
	$Q=4$	62	31.64	36.56	14.08	35.66	−13.32	9.56
SLFM	$Q=2$	34	31.51	24.57	4.44	15.87	−47.66	−45.98
	$Q=3$	51	33.78	41.59	0.88	8.62	−37.25	−24.54
	$Q=4$	68	44.79	43.29	8.93	20.49	−45.63	−37.66
Multi-level ICM	$Q_{ctry}=2,\ Q_{caus}=2$	18	26.42	24.24	4.55	6.84	−48.46	−61.53
	$Q_{ctry}=2,\ Q_{caus}=4$	28	25.74	32.26	5.04	15.75	−35.94	−44.24
	$Q_{ctry}=3,\ Q_{caus}=4$	36	42.03	36.80	3.07	11.12	−44.45	−53.69
Country+Cause+Gender ($L=30$)		#Kernel	# Czech Rep.		Germany		Poland	
		Hyperparameters	APE	CRPS	APE	CRPS	APE	CRPS
ICM	$Q=2$	62	9.99	11.75	−5.40	−13.40	12.21	6.55
	$Q=3$	92	0.52	3.07	−10.78	−20.35	−15.31	4.09
Multi-level ICM	$Q_{ctry}=2,\ Q_{caus}=4,\ Q_{gndr}=2$	32	14.85	24.91	−68.45	−15.22	−25.11	7.75
	$Q_{ctry}=3,\ Q_{caus}=5,\ Q_{gndr}=2$	40	21.34	27.22	−40.45	−10.22	5.12	8.04

(cf. Germany in Table 2). Or it might stabilise inference and improve statistical accuracy, cf. the bottom rows for Poland in Table 2.

Figure 8 shows the predicted log-mortality rates for individual cancers and age group 55–59 years via full-rank multi-level ICM and single-level ICM with $Q=2$. Both models are fitted on age groups 50–84 years and years 1998–2016 before we perform out-of-sample forecasts for the next 3 years (2014–2016). The single-level ICM produces over-smoothed forecasts $m_*(\cdot)$ for several cancers, especially cancers with large observation noise like stomach and pancreas; this problem is mitigated by the shorter lengthscale in year in multi-level ICM ($\theta_{yr} \approx 8.8$ versus $\theta_{yr} \approx 14.8$ in single-level ICM).

Coherence in cause-specific trends: Figure 8 demonstrates that males and females do not always share similar progress in mortality reduction. While the trends in stomach, colorectal and pancreatic cancers are consistent for both genders, for lung cancer the male–female gap is diminishing rapidly, especially in Germany and Poland. This is driven by a decrease in cigarette consumption among men, while women are more likely to develop lung cancers that are not associated with smoking. Thus, the concept of forecast coherence (namely extrapolating a stable male–female spread, as observed historically) is not always well suited for cause-of-death analysis.

5.3 Borrowing the latest datasets from other populations

The period coverage in HCD varies by country as datasets for countries are uploaded asynchronously. This implies that some countries have more up-to-date datasets than others; see Figure A.1 in Appendix A.1 and offers opportunities to fuse data from other countries to update domestic forecasts. For Age-Period-Cohort models, such as Li & Lee (2005), fusion is generally

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

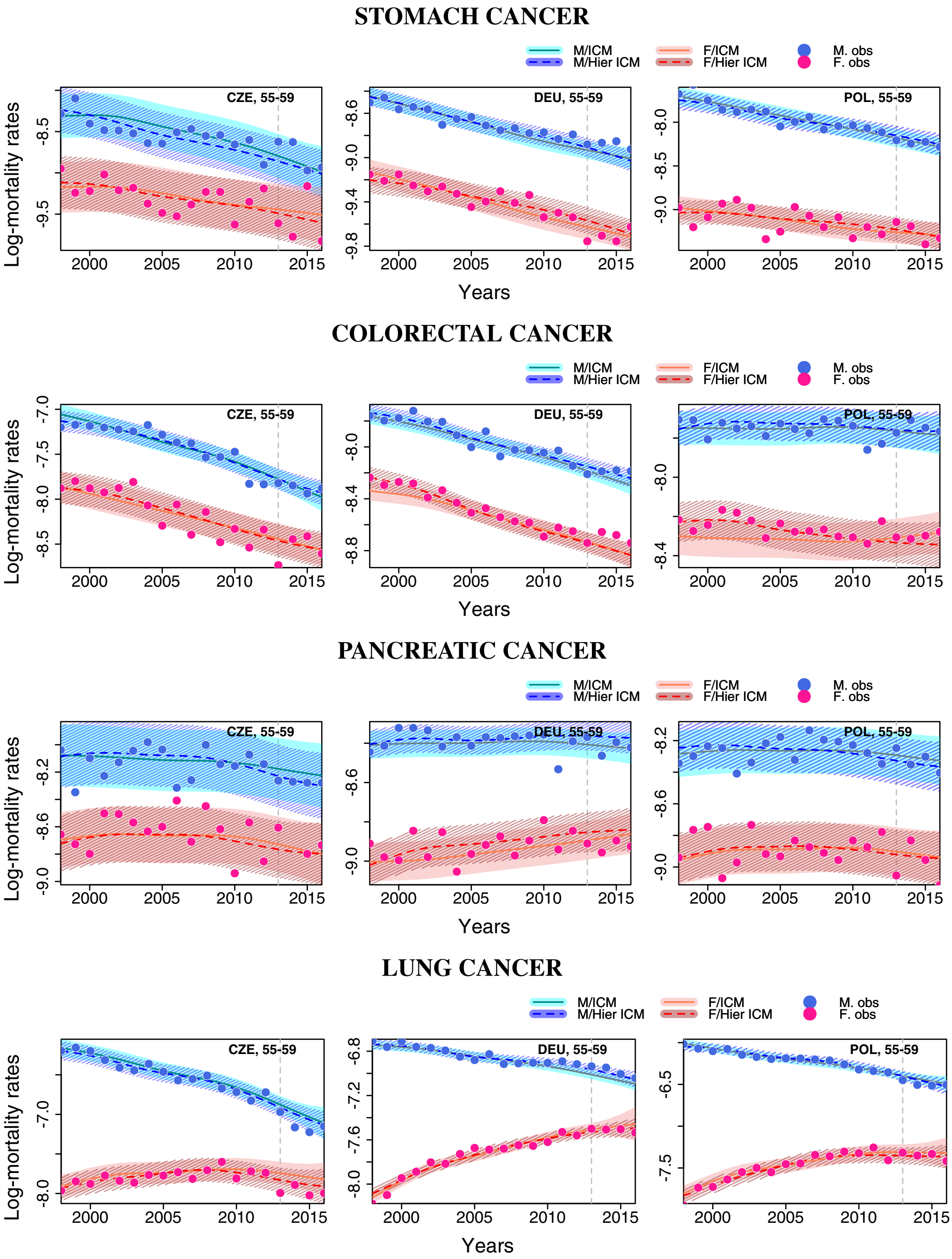


Figure 8. Predictive log-mortality distributions from single-level ICM (listed as ICM, $Q = 2$) and multi-level ICM (listed as Hier ICM, $Q_{ctry} = 3$, $Q_{caus} = 5$, $Q_{gndr} = 2$) models with three factor inputs: country, cause and gender (30 populations). All models are fitted on age groups 50–84 years and years 1998–2013 and applied for 3-year-out forecasts up to 2016. The shadings indicate 95% predictive bands; the vertical lines mark the edge of training data. Note different *y*-axes in each panel.

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

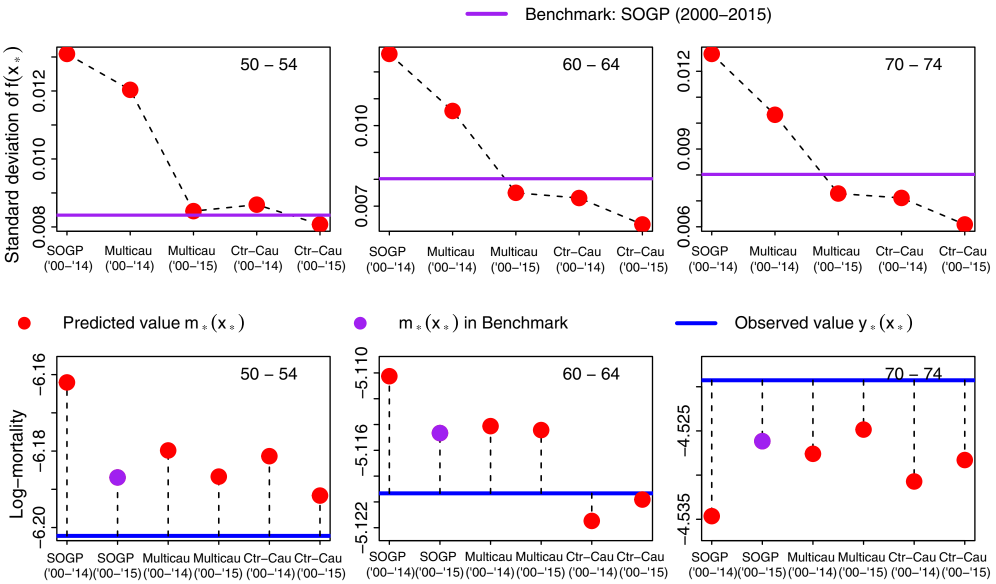


Figure 9. Comparison of the prediction accuracy for 2015 all-cancer log-mortality of French males for indicated age groups between different models with "notched" set-up. Top row: predictive standard deviation $s_*(x_*)$; bottom row: discrepancy between the predictive mean $m_*(x_*)$ and the observed value $y_*(x_*)$.

challenging as model fitting relies on having a rectangular dataset. Our MOGP framework can straightforwardly handle such "notched" datasets to take full advantage of additional observations in different countries.

As an illustration, we consider joint modelling of male mortality in France and Germany. These are two developed Western European countries with similar demographics. Relative to Germany which has data for 1998–2016, France at the time of writing covers only 2000–2015. We choose French males in 2015 as the target population and examine its 1-year-out prediction quality for several models. We take French observations for 2000–2014 and borrow the more up-to-date dataset from Germany (1998–2015) to implement the notched set-up.

We proceed to compare six different models; see Table A.2 in the Appendix. Our comparison covers single-output models for France, multi-cause models for France and multi-cause models joint for France and Germany. Figure 9 visualises relative performance by comparing two components, both for French males in 2015 and across three different age bands: (a) predictive standard deviation $s_*(x_*)$ (top panels) and (b) prediction errors $m_*(x_*) - y_*(x_*)$ relative to the realised 2015 all-cancer log-mortality rates (bottom). Our benchmark is French males SOGP fitted on all-cancer log-mortality rates from 2000 to 2015, that is, the ideal case where latest domestic data are already available. As expected, with access to only 2000–2014 French data, predictions have less credibility (higher $s_*(x_*)$) than the benchmark model that performs in-sample smoothing. However, a country–cause MOGP that ingests both French and German data yields lower $s_*(x_*)$, even without seeing German 2015 mortality. Similarly, multi-cause models boost credibility compared to all-cancer analysis.

For the forecasts, we observe that fusing German data yields prediction errors $|m_*(x_*) - y_*(x_*)|$ that are competitive to that from the benchmark. In fact, the prediction quality of the notched country–cause MOGP (Germany '98-'15, France '00-'14) is as good as the multi-cause MOGP that simultaneously models the log-mortality rates of all five cancer types in French males with 2015 observations available. Throughout, there are the intuitive gains from having 2015 rather than only 2014 experience. In particular, we observe material improvement from access to 2015 German data (right-most two points in Figure 9) which lowers prediction errors.

Remark 5. At the moment, the HCD offers cause-of-death data for less than 15 countries. Many countries do not have recent data available yet (e.g. only up to 2015), leading to limited options in terms of the selection and the number of countries we can incorporate with French males in this experiment. Based on our analysis in Huynh & Ludkovski (2021), choosing countries that are highly correlated with France is essential to maximise the prediction quality. Moreover, the modeller should always double-check the results of a joint model on a validation set before adoption.

6. Conclusion

In this article, we develop multi-output GP models for cause-specific mortality modelling within a multi-population context. With the MOGP mechanism, we are able to capture the cross-cause dependencies that allow joint models to gain more predictive power over single-output models that treat each population independently. Among MOGP variants, SLFM offers more flexibility and is recommended for modelling heterogenous causes, such as top-level ICD categories. Multi-level ICM is demonstrated to work well for interpretable modelling across multiple factor inputs. Our case studies show the applicability of MOGPs to understand cause-specific and aggregate mortality trends, both within a country and across nations, whereby our framework is convenient for information fusion and credibility boosting.

The current work focuses on exploiting the structured Kronecker covariance to efficiently learn the joint covariance kernel. This is sufficient for handling a moderately large number of sub-populations (up to 30 in our case studies); additional techniques would be needed to handle larger datasets, for example, across more causes of death or across all the countries in HCD. There is an active research area looking at alternative methods (local kernel interpolation, inducing points, block structures, etc.; see e.g. Flaxman *et al.*, 2015) for massive scalable GP well suited for gridded mortality datasets. Another methodological extension would be to consider a linear coregionalisation model (LCM) for mortality modelling. LCM generalises ICM and SLFM and allows multiple latent functions from GP priors with different covariance kernels. The third direction would be to investigate other kernel families, such as composite kernels or kernels that can incorporate structural changes. The latter would be useful to model (sub-)causes that exhibit strong disruptions over time in their mortality trends.

Acknowledgement. We thank the anonymous reviewers for constructive suggestions that improved and clarified the final version. Ludkovski is partially supported by NSF-1821240.

References

Alai, D.H., (-Gaille), S. A., Bajekal, M. & Villegas, A. M. (2018). Mind the gap: a study of cause-specific mortality by socioeconomic circumstances. *North American Actuarial Journal*, 22(2), 161–181.

Alvarez, M.A., Rosasco, L. & Lawrence, N.D. (2012). Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, **4**(3), 195–266.

Anderson, R.N., Miniño, A.M., Hoyert, D.L. & Rosenberg, H. M. (2001). Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates. *National Vital Statistics Report*, **49**(2), 1–32.

Arash, E., Ramin, S., Saeid, S. & Sepanlou, S.G. (2020). The global, regional, and national burden of stomach cancer in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet Gastroenterology and Hepatology*, **5**(1), 42–54.

Arnold (-Gaille), S. & Sherris, M. (2013). Forecasting mortality trends allowing for cause-of-death mortality dependence. *North American Actuarial Journal*, 17(4), 273–282.

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

Bergeron-Boucher, **M.-P.**, **Canudas-Romo**, **V.**, **Oeppen**, **J.E.** & **Vaupel**, **J.** (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, **37**(17), 527–566.

Bonilla, **E.V.**, **Chai**, **K.M.** & **Williams**, **C.** (2008). Multi-task Gaussian Process prediction. In *Advances in Neural Information Processing Systems 20* (pp. 153–160). Curran Associates, Inc.

Boumezoued, **A.**, **Coulomb**, **J.-B.**, **Klein**, **A.**, **Louvet**, **D.** & **Titon**, **E.** (2019). Modeling and forecasting cause-of-death mortality, technical report, Society of Actuaries.

Carpenter, **B.**, **Gelman**, **A.**, **D.Hoffman**, **M.**, **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), 1–32.

Caruana, **R.** (1997). Multi-task learning. *Machine Learning*, **28**(1), 41–75.

Caselli, **G.** (1996) Future longevity among elderly populations. In *Health and Mortality among Elderly Populations* (pp. 235–265). Oxford University Press.

Caselli, **G.**, **Vallin**, **J.** & **Marsili**, **M.** (2019). How useful are the causes of death when extrapolating mortality trends. an update. In **T. Bengtsson** & **N. Keilman** (Eds.), *Old and New Perspectives on Mortality Forecasting* (pp. 237–259). Springer International Publishing.

Deville, **Y.**, **Ginsbourger**, **D.**, **Roustant**, **O.** & **Durrande**, **N.** (2019). *kergp: Gaussian Process Laboratory*. R package version 0.5.0.

Dimitrova, **D.S.**, **Haberman**, **S.** & **Kaishev**, **V. K.** (2013). Dependent competing risks: cause elimination and its impact on survival. *Insurance: Mathematics and Economics*, **53**(2), 464–477.

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

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.**, **Wilson**, **A.**, **Neill**, **D.**, **Nickisch**, **H.** & **Smola**, **A.** (2015). Fast Kronecker inference in Gaussian processes with Non-Gaussian likelihoods. In **F. Bach** & **D. Blei** (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 37 (pp. 607–616), Lille, France. PMLR.

Foreman, **K.**, **Marquez**, **N.**, **Dolgert**, **A.**, **Fukutaki**, **K.**, **Fullman**, **N.**, **McGaughey**, **M.**, **Pletcher**, **M.A.**, **Smith**, **A.**, **Tang**, **K.**, **Yuan**, **C.-W.**, **Brown**, **J.**, **Friedman**, **J.**, **He**, **J.**, **Heuton**, **K.**, **Holmberg**, **M.**, **Patel**, **D.J.**, **Reidy**, **P.**, **Carter**, **A.**, **Cercy**, **K.M.**, **Chapin**, **A.**, **Douwes-Schultz**, **D.**, **Frank**, **T.D.**, **Goettsch**, **F.**, **Liu**, **P.**, **Nandakumar**, **V.**, **Reitsma**, **M.**, **Reuter**, **V.**, **Sadat**, **N.**, **Sorensen**, **R.J.D.**, **Srinivasan**, **V.**, **Updike**, **R.**, **York**, **H.**, **Lopez**, **A.D.**, **Lozano**, **R.**, **Lim**, **S.S.**, **Mokdad**, **A.**, **Vollset**, **S.** & **Murray**, **C.** (2018). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *Lancet (London, England)*, **392**, 2052–2090.

Gilboa, **E.**, **Saatçi**, **Y.** & **Cunningham**, **J.P.** (2015). Scaling multidimensional inference for structured Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(2), 424–436.

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

HCD (2021). The Human Cause-of-Death Database. French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany).

Huynh, **N.** & **Ludkovski**, **M.** (2021). Multi-output Gaussian processes for multi-population longevity modelling. *Annals of Actuarial Science*, **15**(2), 318–345.

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.

Kjaergaard, **S.**, **Ergemen**, **Y.E.**, **Kallestrup-Lamb**, **M.**, **Oeppen**, **J.** & **Lindahl-Jacobsen**, **R.** (2019). Forecasting causes of death by using compositional data analysis: the case of cancer deaths. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**(5), 1351–1370.

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

Knudsen, **C.** & **McNown**, **R.** (1993). Changing causes of death and the sex differential in the USA: recent trends and projections. *Population Research and Policy Review*, **12**(1), 27–41.

Letham, **B.** & **Bakshy**, **E.** (2019). Bayesian optimization for policy search via online-offline experimentation. *Journal of Machine Learning Research*, **20**(145), 1–30.

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, **H.** & **Lu**, **Y.** (2019). Modeling cause-of-death mortality using hierarchical Archimedean copula. *Scandinavian Actuarial Journal*, **2019**(3), 247–272.

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

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

Liu, H., Ong, Y.-S., Shen, X. & Cai, J. (2020). When Gaussian process meets big data: a review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, **31**(11), 4405–4423.

Lo, S.M.S. & Wilke, R.A. (2010). A copula model for dependent competing risks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **59**(2), 359–376.

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.

Lyu, P., Waegenaere, A.D. & Melenberg, B. (2021). A multi-population approach to forecasting all-cause mortality using cause-of-death mortality data. *North American Actuarial Journal*, **25**(1), S421–S456.

McNown, R. & Rogers, A. (1992). Forecasting cause-specific mortality using time series methods. *International Journal of Forecasting*, **8**(3), 413–432.

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

Saatçi, Y. (2011). *Sca Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge.

Tabeau, E., Ekamper, P., Huisman, C. & Bosch, A. (1999). Improving overall mortality forecasts by analysing cause-of-death, period and cohort effects in trends. *European Journal of Population/Revue Européenne de Démographie*, **15**(2), 153–183.

Teh, Y., Seeger, M. & Jordan, M. I. (2005). Semiparametric latent factor models. In *International Workshop on Artificial Intelligence and Statistics (AISTATS)* (pp. 333–340). PMLR.

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

WHO (2021). Cancer.

Williams, C., Klanke, S., Vijayakumar, S. & Chai, K. (2009). Multi-task Gaussian process learning of robot inverse dynamics. In **D. Koller, D. Schuurmans, Y. Bengio & Bottou, L.** (Eds.), *Advances in Neural Information Processing Systems*, vol. 21. Curran Associates, Inc.

Wilmoth, J.R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies*, **5**(4), 293–319.

Wojtys, P., Antczak, A. & Godlewski, D. (2014). Predictions of cancer mortality in Poland in 2020. *Central European Journal of Medicine*, **9**, 667–675.

Zhe, S., Xing, W. & Kirby, R.M. (2019). Scalable high-order Gaussian process regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, vol. 89 (pp. 2611–2620). PMLR.

Appendix A. Additional Plots for the Case Study on Cancer Sub-Types

A.1 Year ranges available in the HCD by country

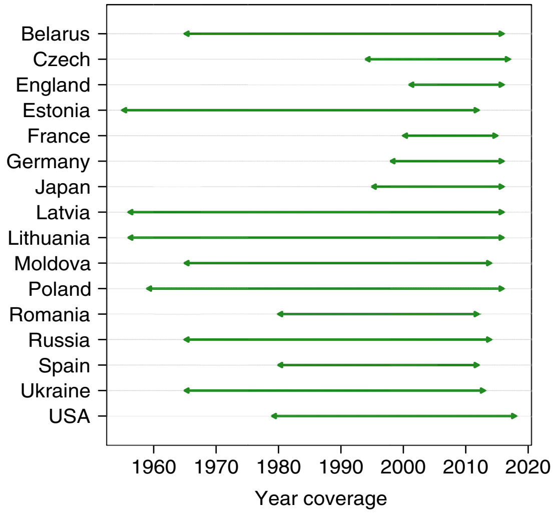


Figure A.1. Countries in the HCD and their historical data coverage as of early 2022.

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

A.2 Age patterns of log-mortality rates of different cancers

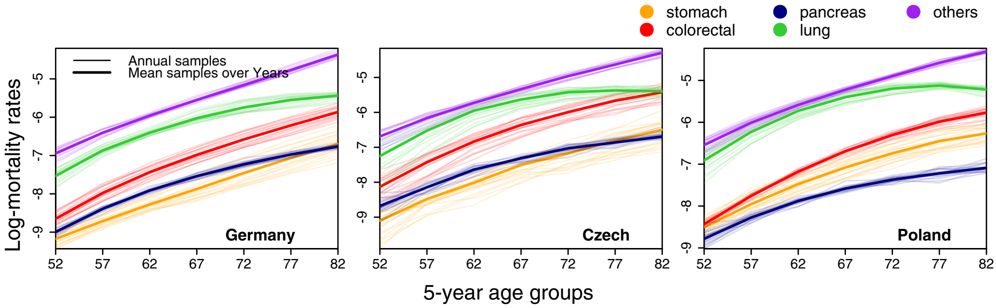


Figure A.2. Raw log mortality as a function of age for five cancer variations in male populations. We show 19 faint curves for each year in the 1998–2016 range used for training, as well as the respective average historical log-mortality rate in bold.

A.3 Kernel hyperparameter learning

Searching for optimal hyperparameters in GP can be challenging if the marginal likelihood features multiple local maxima or flattens around its global maximum (Rasmussen & Williams, 2005). When the optimiser fails to find the global maximum, unsuitable lengthscales in age and year sometimes result, leading to a poor fit of the data. Table A.1 reports the inferred lengthscale in age (θ_{ag}) and year (θ_{yr}) of the SOGP models fitted on the five cancer types for the male populations in the three considered countries. Many SOGP models have θ_{yr} being too short (less than 5 or so), resulting in oscillatory fitted $m_*(\cdot)$'s. Such models lack the ability to distinguish between true signal and the inherent randomness in the data. Similarly, when the estimated lengthscales are too large, the fitted GP surfaces are over-smoothed. Joint models tend to better learn the hyperparameters by enabling data fusion across multiple populations and utilising more observations. In Table A.1, we show that when we fit multi-cause MOGP (both ICM and SLFM) on all five cancer types, the resulting lengthscales are all well calibrated.

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

Table A.1. GP lengthscales θ_{ag}, θ_{yr} in age and year for different models. All models are fitted on age groups 50–84 years and years 1998–2016.

	SOGP on each cancer type					Multi-cause ICM	Multi-cause SLFM	
Czech Rep.	Stomach	Colorectal	Pancreas	Lung	Others	$(Q = 2)$	$(Q = 2)$	
θ_{ag}	159.92	16.32	6.21	17.69	12.52	16.90	21.26	19.67
θ_{yr}	38.49	9.03	7.82	13.87	3.69	10.12	13.05	11.42
Germany	SOGP on each cancer type					Multi-cause ICM	Multi-cause SLFM	
	Stomach	Colorectal	Pancreas	Lung	Others	$(Q = 2)$	$(Q = 2)$	
θ_{ag}	8.65	3.72	8.99	0.00	2.79	8.59	11.75	9.46
θ_{yr}	7.35	4.90	7.69	4.82	4.44	10.26	9.92	6.77
	SOGP on each cancer type					Multi-cause ICM	Multi-cause SLFM	
Poland	Stomach	Colorectal	Pancreas	Lung	Others	$(Q = 2)$	$(Q = 2)$	
θ_{ag}	23.96	16.28	5.72	15.25	22.80	16.98	16.20	21.89
θ_{yr}	253.34	3.69	6.19	9.04	6.01	12.83	11.19	12.11

A.4 Training designs in notched datasets

Table A.2. Descriptions for models being applied to forecast 2015 all-cancer log-mortality rates of French males. All the models are fitted on ages 50–84 years (seven age groups) and year coverage listed below.

GP models	Outcome variable	Country	Year	Pred. type	Abbrev.
SOGP	All-cancer log-mortality	France	2000–2015	In-sample	SOGP (’00-’15)
			2000–2014	Out-of-sample	SOGP (‘00-‘14)
Multi-cause GP	By-cancer log-mortality	France	2000–2015	In-sample	Multi-cause (‘00-‘15)
	(5 variations)		2000–2014	Out-of-sample	Multi-cause (‘00-‘14)
		France	2000–2014	Out-of-sample	Country–cause (‘00-‘14)
	By-cancer log-mortality	Germany	1998–2014		
Country–cause GP					
	(5 variations)	France	2000–2014	Out-of-sample	Country–cause (‘00-‘15)
		Germany	1998–2015		

https://doi.org/10.1017/S1748499523000118 Published online by Cambridge University Press

A.5 Illustrating latent factor loadings in SLFM

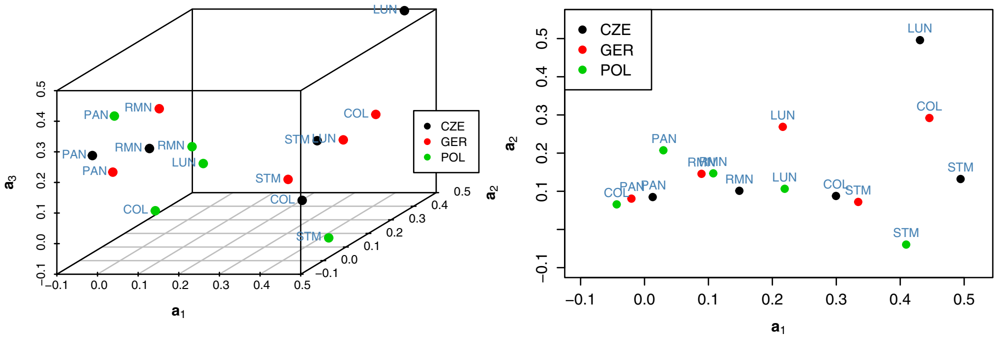
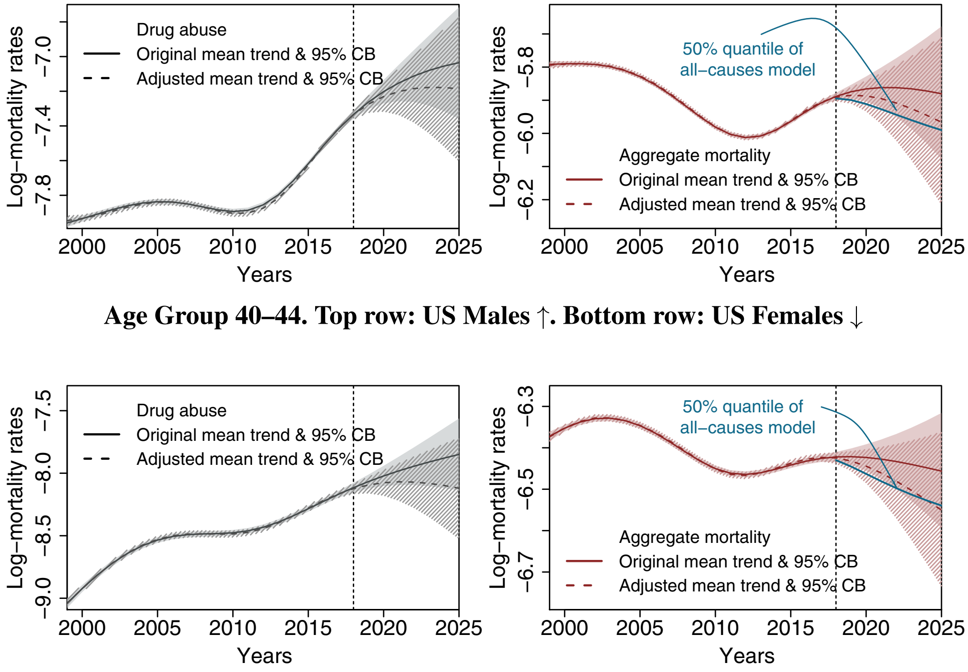


Figure A.3. Factor loadings $\mathbf{a}_q = (a_{1,q}, \ldots, a_{L,q})^T$ in the country–cause SLFM with $Q = 3$ (*left*) and $Q = 2$ (*right*). The model is fitted on ages 50–84 years, years 1998–2016, over three countries and five cancer types. For each of the 15 populations, we plot $(a_{l,1}, a_{l,2}, a_{l,3})$ as a point in three-space on the left panel and $(a_{l,1}, a_{l,2})$ on the right.

Appendix B. Additional Plots for the US Top-Level-Cause Study

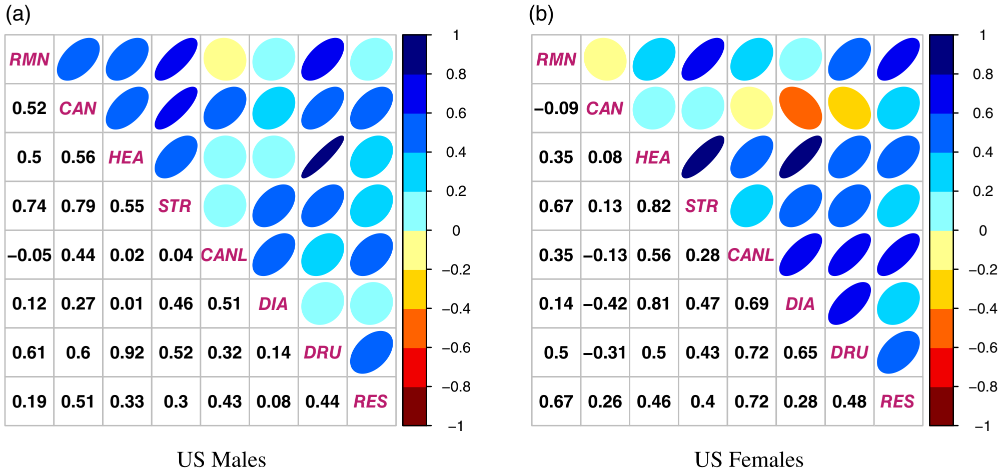
B.1 Adjusting drug overdose trend



Age Group 40–44. Top row: US Males ↑. Bottom row: US Females ↓

Figure B.1. Comparison of original and adjusted predicted log-mortality for age group 40–44 years, US males and Females. Left: same model after reducing drug abuse yearly trend β^{yr} by one-third. Right: corresponding aggregated all-cause trends. Vertical lines indicate the edge of training data (1999–2018).

B.2 Cross-correlation matrices for the US all-cause analysis



RMN							
0.52	CAN						
0.5	0.56	HEA					
0.74	0.79	0.55	STR				
−0.05	0.44	0.02	0.04	CANL			
0.12	0.27	0.01	0.46	0.51	DIA		
0.61	0.6	0.92	0.52	0.32	0.14	DRU	
0.19	0.51	0.33	0.3	0.43	0.08	0.44	RES

US Males

RMN							
−0.09	CAN						
0.35	0.08	HEA					
0.67	0.13	0.82	STR				
0.35	−0.13	0.56	0.28	CANL			
0.14	−0.42	0.81	0.47	0.69	DIA		
0.5	−0.31	0.5	0.43	0.72	0.65	DRU	
0.67	0.26	0.46	0.4	0.72	0.28	0.48	RES

US Females

Figure B.2. Cross-cause correlation matrices derived from multi-cause single-level ($Q = 6$) ICM MOGP, fitted on ages 40–69 years (six age groups) and years 1999–2018, separately for US nales and US females.

Cite this article: Huynh N and Ludkovski M (2024). Joint models for cause-of-death mortality in multiple populations, *Annals of Actuarial Science*, **18**, 51–77. https://doi.org/10.1017/S1748499523000118