

# Is the Group Structure Important in Grouped Functional Time Series?

YANG YANG<sup>1</sup> AND HAN LIN SHANG<sup>2,\*</sup>

<sup>1</sup>*Department of Econometrics and Business Statistics, Monash University, Australia*

<sup>2</sup>*Department of Actuarial Studies and Business Analytics, Level 7, 4 Eastern Road, Macquarie University, Sydney, NSW 2109, Australia*

## Abstract

We study the importance of group structure in grouped functional time series. Due to the non-uniqueness of group structure, we investigate different disaggregation structures in grouped functional time series. We address a practical question on whether or not the group structure can affect forecast accuracy. Using a dynamic multivariate functional time series method, we consider joint modeling and forecasting multiple series. Illustrated by Japanese sub-national age-specific mortality rates from 1975 to 2016, we investigate one- to 15-step-ahead point and interval forecast accuracies for the two group structures.

**Keywords** *dynamic principal component analysis; forecast reconciliation; Japanese sub-national age-specific mortality rates; long-run covariance function; multivariate functional principal component analysis*

## 1 Introduction

Because of the rapid advancements in medical technology and improvements in health services, many countries worldwide have been observing a continuous reduction of human mortality rates for decades. Developed OECD nations, such as Japan, consider the remarkable increase in population life expectancy as one of the greatest achievements of the last century (OECD, 2020). The growing longevity risks in many countries put pressure on government policymakers and planners to develop sustainable pension, health, and aged care systems (see, e.g., Coulmas, 2007). Traditional studies of human mortality focus on population data at the national level. In recent years, academics and government stakeholders have shown increasing interest in regional mortality improvements after realizing that sub-national forecasts of age-specific mortality rates help inform social and economic policies within local regions. Thus, any improvement in the forecast accuracy of mortality rates will be beneficial in determining the allocation of current and future resources at the national and sub-national levels.

Numerous mortality modeling methods have been proposed since the first publication of the Gompertz law in 1825 (see Booth and Tickle, 2008, for a comprehensive literature review on mortality modeling and forecasting). Only a few approaches could simultaneously forecast multiple regional mortality rates within a country (see, e.g., Jarner and Kryger, 2011; Cairns et al., 2011; Dowd et al., 2011; Russolillo et al., 2011; Hatzopoulos and Haberman, 2013; Shang, 2016; Villegas et al., 2017; Gao et al., 2019; Perla et al., 2021). The main hypothesis for multi-

---

\*Corresponding author. Email: [hanlin.shang@mq.edu.au](mailto:hanlin.shang@mq.edu.au).

population mortality models is that mortality rate differences for any two populations having similar socio-economic status and close connections with each other do not diverge indefinitely over time. This paper applies a dynamic multivariate functional time series method to produce accurate sub-national age-specific mortality forecasts.

Applying functional time series forecasting methods to sub-national mortality disaggregated by attributes such as sex and geographical regions independently does not ensure coherence in the forecasts. The forecasts of all sub-populations will not add up to the forecasts obtained by applying the technique to the national data. Hence, in practice, it is crucial to consider forecast reconciliation approaches in forecasting sub-national mortality rates (e.g., see Stone et al., 1942; Sefton and Weale, 2009).

We apply a dynamic multivariate functional time series method to simultaneously model a group of sub-national mortality rates and make forecasts. In particular, we aim to answer a practical issue: does the group structure matter in terms of forecast accuracy for a grouped functional time series? We first arrange the national and sub-national Japanese male and female age-specific mortality rates from 1975 to 2016 into two group structures. Then, we apply dynamic univariate and multivariate functional time series methods and compare their forecasting performance on some holdout data for the two group structures considered.

The rest of this paper is structured as follows. In Section 2, we describe the motivating dataset, which is the Japanese national and sub-national age-specific mortality rates. In Section 3, we describe a dynamic multivariate functional time series method for producing point and interval forecasts. Reconciled forecasts produced by the dynamic univariate and multivariate functional time series forecasting methods are evaluated and compared in Section 4. Conclusions are presented in Section 5, along with some reflections on ways to extend the techniques presented in this study.

## 2 Japanese Age-Specific Mortality Rates

The Japanese Mortality Database regularly recently publishes age-specific mortality rates over the period 1975–2016 at the national level and various subnational levels (Japanese Mortality Database, 2017). We consider ages from 0 to 99 in single years of age and include the last age group, “100+”, as an aggregation of ages at and beyond 100. Figure 1 presents rainbow plots of the age-specific mortality rates of females and males in the studied period. Rainbow plots are designed to display the distant past in red curves and the more recent years in purple curves (Hyndman and Shang, 2010). Typical “J shape” mortality patterns of developed countries can be observed in both panels in Figure 1, which indicates a rapid decrease of mortality occurs for the newborns, with slower reduction continuing into the adolescence age. There is an “accident hump” in mortality for young adults around age 20, followed by a short “plateau” until the early 30s, and then a steady increase in middle-to-old age groups. Figure 1 reveals an overall reduction in the trend of total mortality rates over the considered period 1975–2016. However, the rate of decline appears to vary across different age groups.

### 2.1 Nonparametric Smoothing

We implement a penalized regression spline smoothing method with a monotonic constraint to smooth out excess noise at higher ages (see also Hyndman and Ullah, 2007; Shang, 2016).

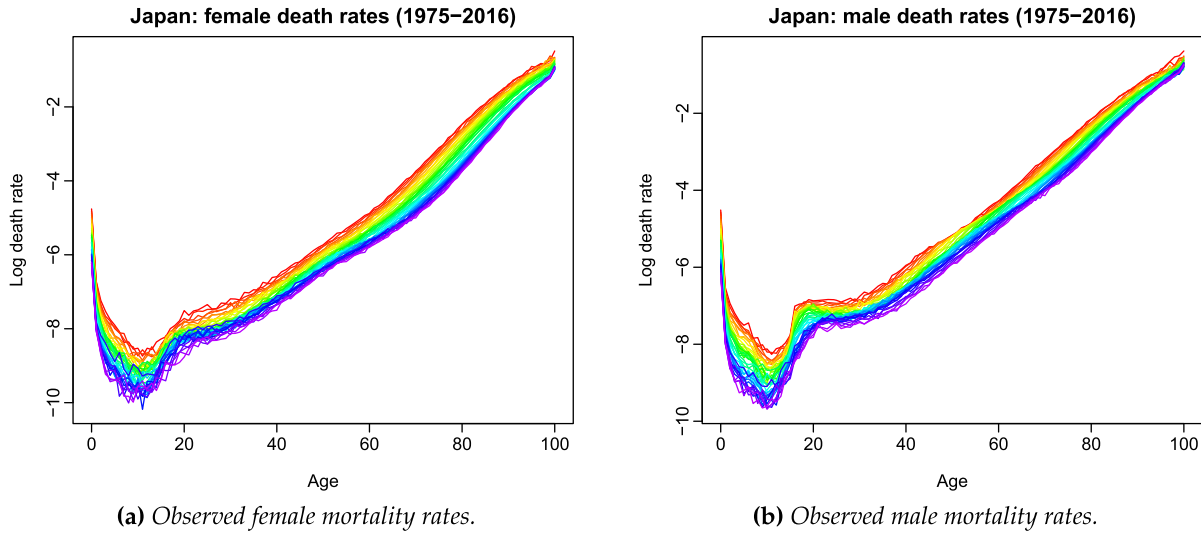


Figure 1: Functional time series graphical displays for the Japanese age-specific mortality rates from 1975 to 2016. Using a rainbow color palette, the mortality rates in the distant past years are shown in red, while the mortality rates in the most recent years are shown in purple.

For a particular population  $j$ , let  $y_t^j(x_i)$  be the log central mortality rates observed at the beginning of each year for year  $t = 1, 2, \dots, n$ , where  $x_i$  represents the center of each age or age group for  $i = 1, \dots, p$ . The value of  $j$  depends on the research question. For example, when modeling age-specific mortality rates of females and males in a particular prefecture,  $j$  takes two values: male or female. We assume there is an underlying continuous and smooth function  $f_t^j(x)$  that is observed at discrete data points with errors of the form

$$y_t^j(x_i) = f_t^j(x_i) + \delta_t^j(x_i)\varepsilon_{t,i}^j,$$

where  $\varepsilon_{t,i}^j$  is an independent and identically distributed standard normal random variable for each age in year  $t$ , and  $\delta_t^j(x_i)$  measures the variability in mortality at each age in year  $t$  for the  $j^{\text{th}}$  population. Jointly,  $\delta_t^j(x_i)\varepsilon_{t,i}^j$  represents the smoothing error.

Let  $m_t^j(x_i) = \exp\{y_t^j(x_i)\}$  be the observed central mortality rates for age  $x_i$  in year  $t$  and define  $E_t^j(x_i)$  to be the total exposure to risks of  $j^{\text{th}}$  population of age  $x_i$  at 1<sup>st</sup> January of year  $t$ . Then  $m_t^j(x_i)$  approximately follows a binomial distribution with estimated variance  $(E_t^j(x_i))^{-1} m_t^j(x_i) [1 - m_t^j(x_i)]$ . Since we are modeling the mortality rates at logarithm scale, we apply a Taylor approximation to  $y_t^j(x) = \ln(m_t^j(x_i))$  to get

$$\left(\widehat{\delta}_t^j\right)^2(x_i) := \text{Var} \left\{ \ln[m_t^j(x_i)] \right\} \approx \frac{1 - m_t^j(x_i)}{m_t^j(x_i) \times E_t^j(x_i)}.$$

Following Hyndman and Ullah (2007), we then define weights  $w_t^j(x_i) = 1/\left(\widehat{\delta}_t^j\right)^2(x_i)$  and use weighted penalized regression splines to obtain the smoothed functions  $f_t^j(x)$  (see also Wood, 2003; Currie et al., 2004).

As a byproduct of the smoothing, we can potentially fill in any missing observations. In Figure 2, we present the smoothed log mortality curves.

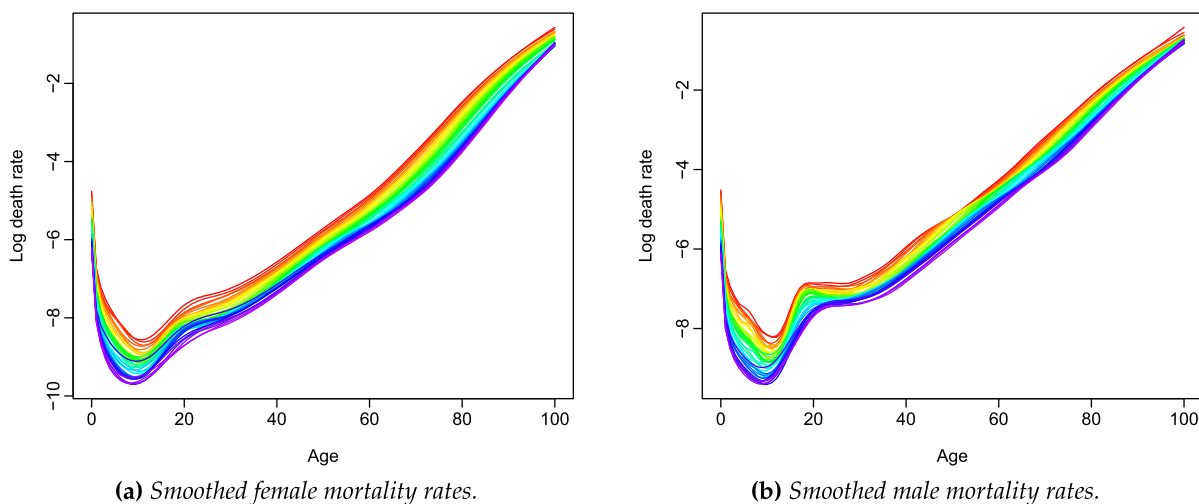


Figure 2: Functional time series graphical displays for the smoothed series, where a penalized regression spline smoothing with monotonic constraint is implemented.

To better understand the mortality patterns of Japan, we consider mortality time series featuring various geographical factors (i.e., country, region, and prefecture) and sex factors (i.e., female and male) in this study. As highlighted by different colors in Figure 3, Japan can be divided into eight geographical regions, namely, Hokkaido (red), Tohoku (yellow), Kanto (green), Chubu (cyan), Kansai (purple), Chugoku (orange), Shikoku (magenta), and Kyushu (grey). Regions vary not only in geography locations, sizes, climates, and terrains but also in histories, traditions, and cultures (The Government of Japan, 2021). Apart from the eight regions, there are 47 prefectures in Japan, forming the country's base level of administrative division. Thus, disaggregate the national and sub-national age-specific mortality rates by geographical location and sex leading to in total 168 series (i.e.,  $47 \times 2 = 94$  Sex  $\times$  Prefecture series, 16 Sex  $\times$  Region series, 2 Sex series at the national level, and further  $1 + 8 + 47 = 56$  total series at Japan, Region and Prefecture levels, respectively). Because of the non-uniqueness of the group structure, the middle-level series can be disaggregated differently. This, in turn, affects the form of a group structure.

We split the Japanese data into geographical structures coupled with a sex grouping variable by considering the difference in mortality between females and males. A three-level geographical structure is used to disaggregate the national and sub-national total mortality rates, as shown in Figure 4(a). We apply a dynamic multivariate functional time series method to eight region-level series and 47 prefecture-level series, respectively, to obtain joint forecasts at each level. Specifically, for each region, we consider all prefecture-level total series that belong to this region; all eight region-level total series are jointly modeled for the national total. These forecasts of multiple series at various levels are then used to perform forecast reconciliation according to the specified group structure.

When forecasting mortality rates of females and males, two disaggregation structures focusing on either geographical factors or sex are used. The first grouping structure (Hierarchy 1) in Figure 4(b) initially disaggregates the top-level series by geographical location. It then further splits sub-national totals into sex-specific series at the region and prefecture levels. Jointly modeling the two sex-specific series at each node of the hierarchy incorporates the close connection



Figure 3: Eight regions (represented in different colors) and 47 prefectures of Japan. In the remaining sections, we mainly use notations  $R_1, \dots, R_8$ , and  $P_1, \dots, P_{47}$  numbered geographically from north to south when referring to regions and prefectures, respectively.

of females and males living in each region or prefecture. The second grouping structure (Hierarchy 2) in Figure 4(c) first splits the total series by sex and then further disaggregates the totals for females and males according to geographical factors. Implementing a dynamic multivariate functional time series method under this structure means jointly modeling the series of one particular sex (female or male at a time) either for each region or for the national sex-specific totals. This sub-national mortality forecasting considers the covariance between spatially closely correlated sub-populations and respects the well-known fact that human females and males have different mortality patterns.

### 3 Methodology

The mortality patterns of sub-populations in Japan are expected to be similar because of their closely related biological backgrounds and similar lifestyles. Still, they are not expected to be identical nationwide across all age groups. Traditional age-specific mortality forecasting methods tend to result in divergent forecasts for sub-populations when independence is assumed. It is desirable to model sub-national mortality rates simultaneously while considering the potential heterogeneity between these. In this paper, we apply a dynamic multivariate functional time series method and demonstrate its advantage over its univariate counterpart using the Japanese sub-national data in Section 4.

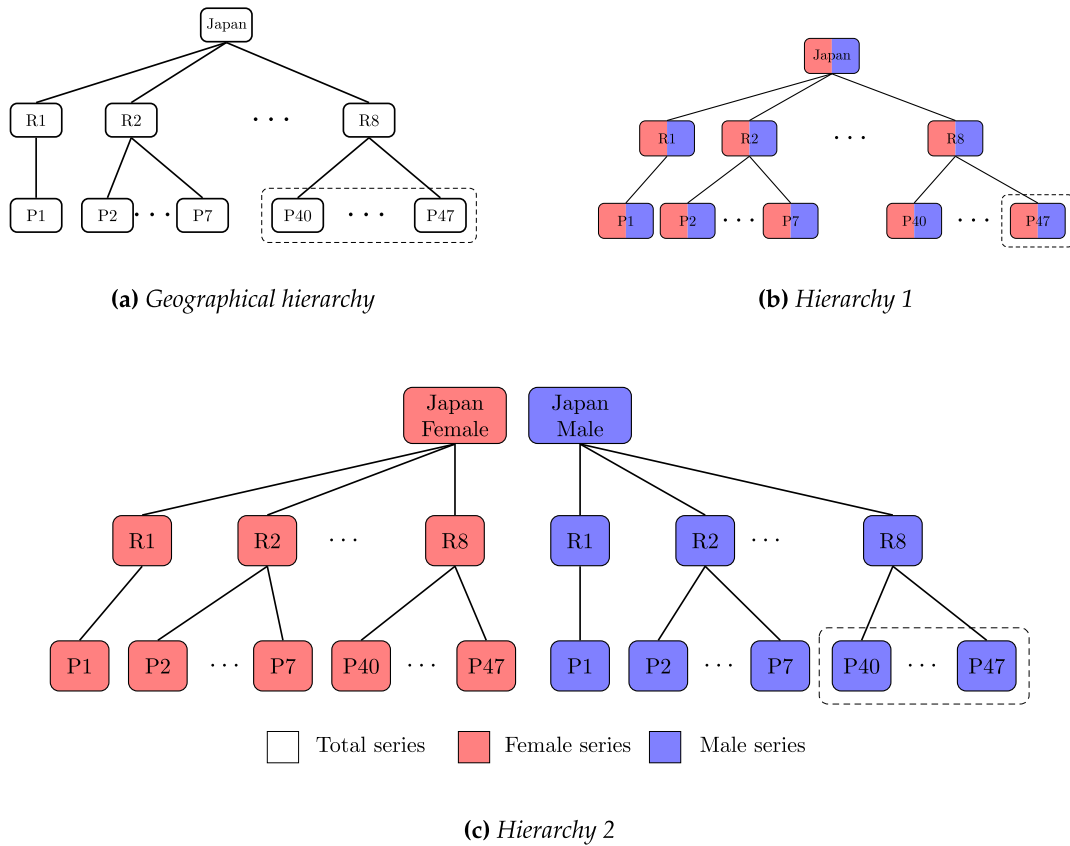


Figure 4: Hierarchy tree diagrams for the Japanese mortality rates. Dashed boxes show examples of jointly modeling multiple series under various hierarchy structures.

### 3.1 Dynamic Multivariate Functional Time Series Method

The dynamic univariate functional time series method considers temporal dependence of data by estimating its long-run covariance function. Let  $\{f^{(l)}(x)\}_{l=1,\dots,\omega}$  be a set of random functions (in total we have  $\omega \in \mathbb{Z}^+$  functions with  $\omega \geq 2$ ) that comprise each observation in the considered data set. Each  $f^{(l)}(x)$  is defined in  $\mathcal{L}^2(\mathcal{I})$ , a Hilbert space of square-integrable functions on an interval  $\mathcal{I} = [a, b]$ ,  $a < b$ . Write  $\mathbf{f}(x) = [f^{(1)}(x), \dots, f^{(\omega)}(x)]^\top$  as a vector in a Hilbert space, consisting of  $\omega$  number of functions. These multivariate functions may be defined on the same domain  $\mathcal{I}$  (see, e.g., Jacques and Preda, 2014; Chiou et al., 2014) or may be defined at different domains  $\mathcal{I}_1, \dots, \mathcal{I}_\omega$  (see, e.g., Happ and Greven, 2018). Given our application has the same domain, we consider the former situation.

We define the mean function

$$\boldsymbol{\mu}(x) := [\mu^{(1)}(x), \dots, \mu^{(\omega)}(x)] = \{E[f^{(1)}(x)], \dots, E[f^{(\omega)}(x)]\}.$$

For  $x, s \in \mathcal{I}$ , auto-covariance function at lag  $\ell$  is defined as

$$\gamma_\ell(x, s) := \text{Cov}[\mathbf{f}_0(x), \mathbf{f}_\ell(s)] = E[(\mathbf{f}_0(x) - \boldsymbol{\mu}(x)), (\mathbf{f}_\ell(s) - \boldsymbol{\mu}(x))],$$

with elements

$$\gamma_\ell^{(ij)}(s, t) := E\left[\left(f_0^{(i)}(x) - E[f^{(i)}(x)]\right)\left(f_\ell^{(j)}(s) - E[f^{(j)}(s)]\right)\right].$$

The long-run covariance is defined as

$$C(x, s) := \sum_{\ell=-\infty}^{\infty} \gamma_{\ell}(x, s), \tag{1}$$

where  $f^{(j)}(s)$  and  $f^{(l)}(x) \in \mathcal{L}^2(\mathcal{I})$ . The long-run covariance function  $C(x, s)$  consists of the variance as well as the temporal cross-covariance at various positive and negative lags of a functional time series, and is a wall-defined element of  $\mathcal{L}^2(\mathcal{I})$  under mild dependence and moment conditions. Through right integration,  $C(x, s)$  defines a Hilbert-Schmidt integral operator on  $\mathcal{L}^2(\mathcal{I})$  given by

$$(\mathcal{K}(f))^{(j)}(s) = \sum_{l=1}^{\omega} \int_{\mathcal{I}} C_{lj}(x, s) f^{(l)}(x) dx,$$

where  $\mathcal{K}$  denotes a kernel operator  $\mathcal{L}^2(\mathcal{I}) \rightarrow \mathcal{L}^2(\mathcal{I})$  and  $C_{lj}(x, s)$  is the  $(l, j)^{\text{th}}$  element of  $C(x, s)$ .

Via Mercer’s lemma, there exists an orthonormal sequence  $(\phi_k)$  of continuous functions in  $\mathcal{L}^2(\mathcal{I})$  and a non-increasing sequence  $\lambda_k$  of positive number, such that

$$C_{lj}(x, s) = \sum_{k=1}^{\infty} \lambda_k \phi_k^{(l)}(x) \phi_k^{(j)}(s).$$

By the separability of Hilbert spaces, the Karhunen-Loève expansion of a stochastic process  $f^{(l)}(x) = \{f_1^{(l)}(x), \dots, f_n^{(l)}(x)\}$  can be expressed as

$$\begin{aligned} f_t^{(l)}(x) &= \mu^{(l)}(x) + \sum_{k=1}^{\infty} \beta_{t,k}^{(l)} \phi_k^{(l)}(x) \\ &= \mu^{(l)}(x) + \sum_{k=1}^K \beta_{t,k}^{(l)} \phi_k^{(l)}(x) + e_t^{(l)}(x), \end{aligned} \tag{2}$$

where  $\mu^{(l)}(x)$  is the mean function for the  $l^{\text{th}}$  sub-population;  $\{\phi_1^{(l)}(x), \dots, \phi_K^{(l)}(x)\}$  is a set of the first  $K$  functional principal components;  $\beta_1^{(l)} = (\beta_{1,1}^{(l)}, \dots, \beta_{n,1}^{(l)})^{\top}$  and  $\{\beta_1^{(l)}, \dots, \beta_K^{(l)}\}$  denotes a set of principal component scores for the  $l^{\text{th}}$  sub-population in (1);  $e_t^{(l)}(x)$  denotes the model truncation error function with mean zero and finite variance for the  $l^{\text{th}}$  sub-population; and  $K < n$  is the number of retained principal components. Expansion (2) facilitates dimension reduction as the first  $K$  terms often provide a good approximation to the infinite sums, and thus the information contained in  $f^{(l)}(x)$  can be adequately summarized by the  $K$ -dimensional vector  $(\beta_1^{(l)}, \dots, \beta_K^{(l)})$ . Truncating at the first  $K^{\text{th}}$  functional principal components gives the approximation in its matrix formulation as

$$f_t(x) = \Phi \beta_t^{\top},$$

where  $f_t(x) = [f_t^{(1)}(x), \dots, f_t^{(\omega)}(x)]$ ,  $\beta_t = (\beta_{t,1}^{(1)}, \dots, \beta_{t,K}^{(1)}, \beta_{t,1}^{(2)}, \dots, \beta_{t,K}^{(2)}, \dots, \beta_{t,1}^{(\omega)}, \dots, \beta_{t,K}^{(\omega)})$  being the vector of the basis expansion coefficients, and

$$\Phi(x) = \begin{pmatrix} \phi_1^{(1)}(x) & \dots & \phi_K^{(1)}(x) & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \phi_1^{(2)}(x) & \dots & \phi_K^{(2)}(x) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \phi_1^{(\omega)}(x) & \dots & \phi_K^{(\omega)}(x) \end{pmatrix}_{\omega \times (K \times \omega)}.$$

The dynamic functional principal component analysis is an effective way of extracting a large amount of long-run covariance. It summarizes the main features of an infinite-dimensional object by its finite elements by minimizing the mean integrated squared error of the reconstruction error function over the whole functional data set. The dynamic functional principal component analysis has been applied in modeling human age-specific mortality (Shang, 2019) and foreign exchange rates (Shang and Kearney, 2021). This work is the first that extends the technique to model multivariate functional time series to the best of our knowledge. The dynamic multivariate functional time series method could better capture the temporal dependency of age-specific mortality rates compared to the commonly used Lee-Carter model or static functional time series method (see, e.g., Hyndman and Ullah, 2007).

### 3.2 Long-Run Covariance Estimation

The long-run covariance function introduced in (1) in practice can be estimated from a finite sample  $\mathbf{f}(x) = \{\mathbf{f}_1(x), \dots, \mathbf{f}_n(x)\}$ . A natural estimator of  $C(x, s)$  in a bi-infinite sum is given by

$$\widehat{C}_{v,q}(x, s) = \sum_{\ell=-\infty}^{\infty} W_q\left(\frac{\ell}{v}\right) \widehat{\gamma}_{\ell}(x, s), \quad (3)$$

where  $v$  is the bandwidth parameter, and

$$\widehat{\gamma}_{\ell}(x, s) = \begin{cases} \frac{1}{n} \sum_{t=1}^{n-\ell} [\mathbf{f}_t(x) - \widehat{\boldsymbol{\mu}}(x)][\mathbf{f}_{t+\ell}(s) - \widehat{\boldsymbol{\mu}}(s)], & \ell \geq 0; \\ \frac{1}{n} \sum_{t=1-\ell}^n [\mathbf{f}_t(x) - \widehat{\boldsymbol{\mu}}(x)][\mathbf{f}_{t+\ell}(s) - \widehat{\boldsymbol{\mu}}(s)], & \ell < 0, \end{cases}$$

with  $\widehat{\boldsymbol{\mu}}(x) = n^{-1} \sum_{t=1}^n \mathbf{f}_t(x)$ . The weight function  $W_q$  in (3) is a continuous and symmetric function with bounded support of order  $q$  defined on  $[-m, m]$  for some  $m > 0$  satisfying

$$W_q(0) = 1, W_q(u) \leq 1, W_q(u) = W_q(-u), W_q(u) = 0 \text{ if } |u| > m,$$

and there exists an  $\omega$  such that

$$0 < \omega = \lim_{u \rightarrow 0} |u|^{-q} (1 - W_q(u)) < \infty.$$

As with the kernel sandwich estimator in (3), the crucial part is the estimation of bandwidth parameter  $v$ . We estimated  $v$  through a data-driven approach using the plug-in algorithm of Rice and Shang (2017, Section 2). The plug-in algorithm selects the optimal bandwidth  $v$  that minimizes the asymptotic mean-squared normed error between the estimated and actual long-run variance functions.

### 3.3 Selecting the Number of Retained Functional Principal Components

There are several methods for selecting the number of functional principal components  $K$ : (1) the fraction of the variance explained by the first few functional principal components (Chiou, 2012); (2) pseudo-versions of the Akaike information criterion (AIC) and Bayesian information criterion (Yao et al., 2005); (3) predictive cross-validation leaving one or more curves out (Rice and Silverman, 1991); and (4) bootstrap methods (Hall and Vial, 2006). Here, the number of



components is determined as the minimum that reaches 90% of total variance explained by the leading components (the first method mentioned), such that

$$K = \operatorname{argmin}_{K:K \geq 1} \left\{ \sum_{k=1}^K \widehat{\lambda}_k / \sum_{k=1}^{\infty} \widehat{\lambda}_k \mathbb{1}_{\{\widehat{\lambda}_k > 0\}} \geq 0.9 \right\},$$

where  $\mathbb{1}_{\{\widehat{\lambda}_k > 0\}}$  is to exclude possible zero eigenvalues, and  $\mathbb{1}\{\cdot\}$  represents the binary indicator function. Although the total variance explained level could be arbitrary, 0.9 is a common threshold when using the total variance explained in principal component analysis to select the number of components to retain (see also Kolkiewicz et al., 2021).

### 3.4 Forecasting Multivariate Functional Time Series

In the dynamic multivariate functional time series method, we first stack all series into a vector of functions and obtain a set of dynamic functional principal components with associated scores according to the method described in Section 3.1. Conditioning on the estimated mean function and estimated dynamic functional principal components, these forecast scores are multiplied by the estimated principal components to form forecast curves. A univariate time series forecasting method, such as the ARIMA model, is considered to forecast dynamic functional principal component scores. Through differencing, the ARIMA model can handle non-stationarity present in the data. For details of forecasting dynamic principal component scores, consult Hyndman and Shang (2009), Shang (2019) and Shang and Kearney (2021).

Denote the  $h$ -step-ahead prediction of dynamic principal component scores  $\{\beta_1^{(l)}, \dots, \beta_K^{(l)}\}$  by  $\{\widehat{\beta}_{n+h,1}^{(l)}, \dots, \widehat{\beta}_{n+h,K}^{(l)}\}$ . We construct the  $h$ -step-ahead point forecast of  $f_{n+h}(x)$  as

$$\widehat{f}_{n+h}^{(l)}(x) = \mathbb{E} \left[ f_{n+h}^{(l)}(x) | \mathbf{f}^{(l)}(x), \mathbf{B} \right] = \widehat{\mu}^{(l)}(x) + \sum_{k=1}^K \widehat{\beta}_{n+h,k}^{(l)} \widehat{\phi}_k^{(l)}(x),$$

where  $\mathbf{B} = \{\widehat{\phi}_1^{(l)}(x), \dots, \widehat{\phi}_K^{(l)}(x)\}$  are the estimated functional principal component functions.

### 3.5 Construction of Pointwise Prediction Intervals

As a supplement to point forecast evaluation, prediction intervals are constructed and evaluated to gauge the uncertainty associated with point forecasts. Depending on statistical theory and data on error distributions, prediction intervals explicitly estimate the probability that future realizations lie within a given range. In implementing the proposed mortality forecasting method, the errors accumulated in the estimated principal component scores and model residuals in (2) are the primary sources of uncertainty. As emphasized by Chatfield (1993), it is crucial to provide interval forecasts as well as point forecasts to be able to (1) assess future uncertainty level; (2) enable different strategies to be planned for the range of possible outcomes; (3) compare forecasts from different methods; and (4) explore different scenarios based on various underlying assumptions.

To construct uniform and pointwise prediction intervals, we adopt the method of Aue et al. (2015). The method can be summarized in the following steps:

- 1) Using all observed data, compute the  $K$ -variate score vectors  $(\boldsymbol{\beta}_1^{(l)}, \dots, \boldsymbol{\beta}_K^{(l)})$  and the sample functional principal components  $[\widehat{\phi}_1^{(l)}(x), \dots, \widehat{\phi}_K^{(l)}(x)]$ . Then, we can construct in-sample forecasts

$$\widehat{f}_{\xi+h}^{(l)}(x) = \widehat{\beta}_{\xi+h,1}^{(l)} \widehat{\phi}_1^{(l)}(x) + \dots + \widehat{\beta}_{\xi+h,K}^{(l)} \widehat{\phi}_K^{(l)}(x),$$

where  $(\widehat{\beta}_{\xi+h,1}^{(l)}, \dots, \widehat{\beta}_{\xi+h,K}^{(l)})$  are the elements of the  $h$ -step-ahead prediction obtained from  $(\boldsymbol{\beta}_1^{(l)}, \dots, \boldsymbol{\beta}_K^{(l)})$  by a univariate time series forecasting method, for  $\xi \in \{K, K+1, \dots, n-h\}$ .

- 2) With the in-sample point forecasts, we calculate the in-sample point forecast errors

$$\widehat{\epsilon}_{\zeta}^{(l)}(x) = f_{\xi+h}^{(l)}(x) - \widehat{f}_{\xi+h}^{(l)}(x),$$

where  $\zeta \in \{1, 2, \dots, M\}$  and  $M = n - h - K + 1$ .

- 3) Based on these in-sample forecast errors, we can sample with replacement to obtain a series of bootstrapped forecast errors, from which we obtain lower and upper bounds, denoted by  $\gamma_{\text{lb}}^{(l)}(x)$  and  $\gamma_{\text{ub}}^{(l)}(x)$ , respectively. We then seek a tuning parameter  $\pi_{\alpha}$  such that  $\alpha \times 100\%$  of the residuals satisfy

$$\pi_{\alpha} \times \gamma_{\text{lb}}^{(l)}(x_i) \leq \widehat{\epsilon}_{\zeta}^{(l)}(x_i) \leq \pi_{\alpha} \times \gamma_{\text{ub}}^{(l)}(x_i), \quad i = 1, \dots, p.$$

Then, the  $h$ -step-ahead pointwise prediction intervals are given as

$$\pi_{\alpha} \times \gamma_{\text{lb}}^{(l)}(x_i) \leq f_{n+h}^{(l)}(x_i) - \widehat{f}_{n+h}^{(l)}(x_i) \leq \pi_{\alpha} \times \gamma_{\text{ub}}^{(l)}(x_i),$$

where  $i$  symbolizes the discretized data points.

Note that Aue et al. (2015) calculate the standard deviation of  $[\widehat{\epsilon}_1^{(l)}(x), \dots, \widehat{\epsilon}_M^{(l)}(x)]$ , which leads to a parametric approach to constructing prediction intervals. Instead, we consider the nonparametric bootstrap approach of Shang (2018), which allows us to reconcile bootstrapped forecasts among different functional time series in a hierarchy.

### 3.6 Reconciliation of Grouped Forecasts

The disaggregated data can first be organized into groups for forecasting sub-national mortality rates before applying the dynamic multivariate functional time series method. Hence, an appropriate grouping structure is crucial for accurate forecasts. A natural strategy is arranging the sub-national functional time series into geographical structures, as in Figure 4. State-of-the-art forecast reconciliation methods are applied to ensure that the point and interval forecasts at the various disaggregation levels add up to the forecasts of the national mortality.

For all aforementioned hierarchies, we denote a particular disaggregated series using the notation  $G * S$ , meaning the geographical area  $G$  and the sex  $S$ . For instance,  $R_1 * F$  denotes females in Region 1,  $P_1 * T$  denotes females and males in Prefecture 1 and  $Japan * M$  denotes males in Japan. Let  $E_{G*S,t}(x)$  denote the exposure to risk for series  $G * S$  in year  $t$  and age  $x$ , and let  $D_{G*S,t}(x)$  be the number of deaths for series  $G * S$  in year  $t$  and age  $x$ . Then, age-specific

mortality rate is given by  $R_{G*S,t}(x) = D_{G*S,t}(x)/E_{G*S,t}(x)$ . Dropping the age argument ( $x$ ) allows us to express the national and sub-national mortality series in a matrix multiplication as

$$\underbrace{\begin{bmatrix} R_{Japan*T,t} \\ R_{Japan*F,t} \\ R_{Japan*M,t} \\ R_{R1*T,t} \\ R_{R2*T,t} \\ \vdots \\ R_{R8*T,t} \\ R_{R1*F,t} \\ R_{R2*F,t} \\ \vdots \\ R_{R8*F,t} \\ R_{R1*M,t} \\ R_{R2*M,t} \\ \vdots \\ R_{R8*M,t} \\ R_{P1*T,t} \\ R_{P2*T,t} \\ \vdots \\ R_{P47*T,t} \\ R_{P1*F,t} \\ R_{P1*M,t} \\ R_{P2*F,t} \\ R_{P2*M,t} \\ \vdots \\ R_{P47*F,t} \\ R_{P47*M,t} \end{bmatrix}}_{\mathbf{R}_t} = \begin{bmatrix} \frac{EP1*F,t}{E_{Japan*T,t}} & \frac{EP1*M,t}{E_{Japan*T,t}} & \frac{EP2*F,t}{E_{Japan*T,t}} & \frac{EP2*M,t}{E_{Japan*T,t}} & \frac{EP3*F,t}{E_{Japan*T,t}} & \frac{EP3*M,t}{E_{Japan*T,t}} & \dots & \frac{EP47*F,t}{E_{Japan*T,t}} & \frac{EP47*M,t}{E_{Japan*T,t}} \\ \frac{EP1*F,t}{E_{Japan*F,t}} & 0 & \frac{EP2*F,t}{E_{Japan*F,t}} & 0 & \frac{EP3*F,t}{E_{Japan*F,t}} & 0 & \dots & \frac{EP47*F,t}{E_{Japan*F,t}} & 0 \\ 0 & \frac{EP1*M,t}{E_{Japan*M,t}} & 0 & \frac{EP2*M,t}{E_{Japan*M,t}} & 0 & \frac{EP3*M,t}{E_{Japan*M,t}} & \dots & 0 & \frac{EP47*M,t}{E_{Japan*M,t}} \\ \frac{EP1*F,t}{E_{R1*T,t}} & \frac{EP1*M,t}{E_{R1*T,t}} & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \frac{EP2*F,t}{E_{R2*T,t}} & \frac{EP2*M,t}{E_{R2*T,t}} & \frac{EP3*F,t}{E_{R2*T,t}} & \frac{EP3*M,t}{E_{R2*T,t}} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \frac{EP47*F,t}{E_{R8*T,t}} & \frac{EP47*M,t}{E_{R8*T,t}} \\ \frac{EP1*F,t}{E_{R1*F,t}} & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \frac{EP2*F,t}{E_{R2*F,t}} & 0 & \frac{EP3*F,t}{E_{R2*F,t}} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \frac{EP47*F,t}{E_{R8*F,t}} & 0 \\ 0 & \frac{EP1*M,t}{E_{R1*M,t}} & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \frac{EP2*M,t}{E_{R2*M,t}} & 0 & \frac{EP3*M,t}{E_{R2*M,t}} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & \frac{EP47*M,t}{E_{R8*M,t}} \\ \frac{EP1*F,t}{E_{P1*T,t}} & \frac{EP1*M,t}{E_{P1*T,t}} & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \frac{EP2*F,t}{E_{P2*T,t}} & \frac{EP2*M,t}{E_{P2*T,t}} & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \frac{EP47*F,t}{E_{P47*T,t}} & \frac{EP47*M,t}{E_{P47*T,t}} \\ 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}}_{\mathbf{S}_t}$$

$$\times \underbrace{\begin{bmatrix} R_{P1*F,t} \\ R_{P1*M,t} \\ R_{P2*F,t} \\ R_{P2*M,t} \\ \vdots \\ R_{P47*F,t} \\ R_{P47*M,t} \end{bmatrix}}_{\mathbf{b}_t}$$

or  $\mathbf{R}_t = \mathbf{S}_t \mathbf{b}_t$ , where  $\mathbf{R}_t$  is a vector containing all series at all levels of disaggregation,  $\mathbf{b}_t$  is a vector of the most disaggregated series, and  $\mathbf{S}_t$  shows how the two are related. The group

structures illustrated in Figure 4 satisfy this relationship. In the two group structures, the main difference is about how the  $(R_{R1 \times F, t}, R_{R1 \times M, t}, \dots, R_{R8 \times F, t}, R_{R8 \times M, t})$  and  $(R_{P1 \times F, t}, R_{P1 \times M, t}, \dots, R_{P47 \times F, t}, R_{P47 \times M, t})$  are estimated. In Figure 4(b), we estimate the female and male series together in a geographical location. In Figure 4(c), we estimate multiple series at various geographical locations for females and males, respectively.

We present a brief review of three methods for forecast reconciliation based on this summing matrix. The three forecast reconciliation methods are bottom-up, optimal combination method with the ordinary least squares (OLS) estimator, and optimal combination method with generalized least squares (GLS) estimator. The bottom-up method has the agreeable feature that it is intuitive and straightforward and always results in forecasts that satisfy the same group structure as the original data (e.g., see Dangerfield and Morris, 1992; Zellner and Tobias, 2000). Instead of considering only time series in the bottom-level, Hyndman et al. (2011) proposed a method in which base forecasts for all aggregated and disaggregated series are computed successively. Then, the resulting forecasts are combined through linear regression. With the OLS estimator, the reconciled forecasts are close to the base forecasts and aggregate consistently within the group. Wickramasuriya et al. (2019) proposed an improved approach of finding coherent forecasts for the grouped functional time series. By minimizing the sum of variances of reconciliation errors, essentially minimizing trace (MinT) of the variance matrix, this approach produces coherent forecasts across the entire collection of time series. Since the variance-covariance matrix of out-of-sample reconciliation errors is often unknown and not identifiable in practice, the MinT method attempts to approximate weights with the in-sample base forecast errors using the GLS estimator.

## 4 Data Analysis Results

We apply the dynamic univariate and multivariate functional time series forecasting methods to Japanese age-specific mortality rates to obtain base forecasts for the group structures in Figure 4. Then, we conduct reconciliation via the bottom-up and the optimal combination methods. To assess model and parameter stabilities over time, we consider an expanding window analysis of considered time series models (see Zivot and Wang, 2006, Chapter 9 for details). We used the observed mortality curves in 1975–2001 Japanese data to produce one- to 15-step-ahead point and interval forecasts. Through an expanding window approach, we re-estimate the parameters of considered models using the first 28 years of observations from 1975 to 2002 and estimate one- to 14-step-ahead forecasts using re-estimated models. The process is iterated with the sample size increased by one year each time until reaching the end of the data period in 2016. This process produces 15 one-step-ahead forecasts, 14 two-step-ahead forecasts, and so on, up to one 15-step-ahead forecast. We evaluate the point and interval forecast accuracies and report our study results in this section.

### 4.1 Point Forecast Evaluation and Comparison

To evaluate the point forecast accuracy, we compute the root mean squared forecast error (RMSFE), which measures how close the forecasts are compared to the actual values. For each

series, the RMSFE is given by

$$\text{RMSFE}(h) = \sqrt{\frac{1}{101 \times (16 - h)} \sum_{\zeta=h}^{15} \sum_{i=1}^{101} [f_{n+\zeta}(x_i) - \hat{f}_{n+\zeta}(x_i)]^2},$$

where  $f_{n+\zeta}(x_i)$  represents the actual holdout sample for the  $i^{\text{th}}$  age and  $\zeta^{\text{th}}$  curve of the forecasting period, and  $\hat{f}_{n+\zeta}(x_i)$  is the corresponding point forecasts. We obtain measures of point forecast accuracies for all all national and sub-national mortality series, averaging over 15 forecast horizons, as

$$\text{Mean (RMSFE)} = \frac{1}{15} \sum_{h=1}^{15} \text{RMSFE}(h).$$

Averaging all series at each disaggregation level of Hierarchy 1, we show RMSFEs ( $\times 100$ ) for the forecasts before and after reconciliations in Figure 5. In the figure, we denote the independent forecasts as Base. We denote the reconciled forecasts obtained by the bottom-up method and the optimal combination method as BU and OP, respectively (see, e.g., Hyndman et al., 2011; Shang, 2017, for more details). This figure compares forecasts obtained by the dynamic multivariate functional time series method with forecasts produced by its univariate counterpart method. The dynamic multivariate functional time series method is expected to give more accurate point forecasts at disaggregation levels where multiple mortality series have highly consistent patterns. The correlations among multiple series can be captured and used in producing forecasts. This advantage of our method is shown in “Region + Sex series” and “Prefecture + Sex series” panels of Figure 5. However, due to its construction, the dynamic multivariate functional time series method implicitly assumes the same number of retained functional principal components for all series, consequently resulting in diminishing forecast accuracy for a particular series. This disadvantage of our method is revealed when only the Japan \* F and Japan \* M series are considered, as shown in the “Sex series” panel of Figure 5.

At the national level, where only one mortality time series exists, the dynamic univariate functional time series method, a special case of dynamic multivariate functional time series method, is implemented based on an estimated long-run covariance function. Hence, both methods have the same forecasts for the total Japanese series.

The advantage of dynamic multivariate functional time series method is that it could extract common features among strongly related sub-populations. Another prominent character of the dynamic multivariate functional time series method shown in Figure 5 is the method’s robustness to outliers in the training sample. At least three prefectures in Japan that were severely affected by earthquakes and tsunamis have inflated mortality curves. One reason for the dynamic multivariate functional time series forecasts being immune to outliers is that the computation of long-run covariance function over a period mitigates the influence of abrupt and temporal movements of observations. The autocovariance function, a part of the long-run covariance function, can smooth out the observational noise. Another reason for the robust forecast is that collectively, the consideration of related sub-populations enables the information pooling and thus improves the extraction of human mortality features. By contrast, many large values exist in the dynamic univariate functional time series forecasts even after reconciliation.

Figure 5 also shows that pairing all three reconciliation methods with either dynamic multivariate or univariate functional time series method improves the point forecast accuracy. Among

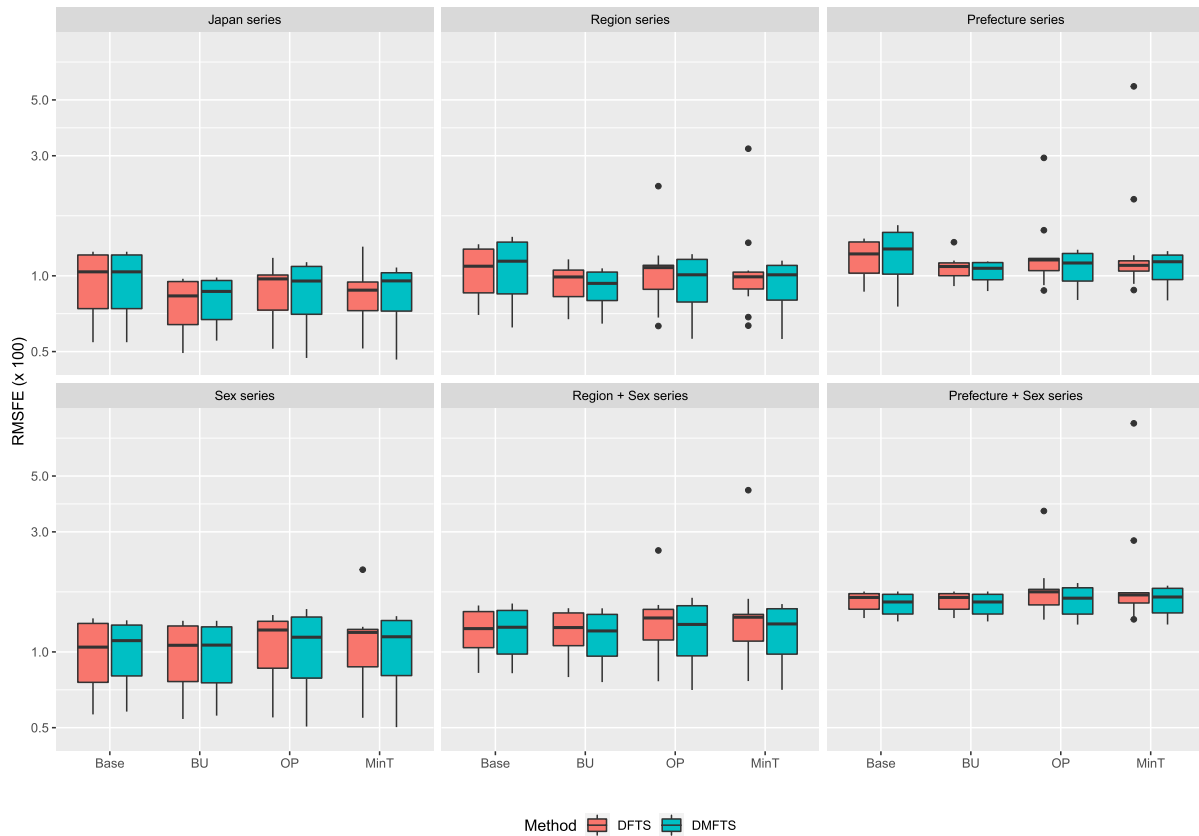


Figure 5: Point forecasting performance of the independent (Base) and reconciled forecasts (BU, OP, and MinT) for univariate and multivariate functional time series methods. DFTS denotes the dynamic univariate functional time series method, while DMFTS denotes the dynamic multivariate functional time series method.

the reconciliation methods considered, the bottom-up method yields the most accurate overall point forecasts when used with dynamic multivariate functional time series, owing to the high signal-to-noise ratio of the prefecture-level sex-specific mortality series. Our results confirm Shang and Hyndman's (2017) early finding regarding the superiority of the bottom-up method for the Japanese sub-national age-specific mortality rates.

Applying the dynamic multivariate functional time series method to various disaggregation structures leads to different RMSFEs ( $\times 100$ ), as shown in Figure 6. The geographical hierarchies described in Figure 4 are blended into Hierarchy 1 and Hierarchy 2 in the reconciliation of the total series at the country, region, and prefecture levels. Hierarchy 2 produces the most accurate point forecasts, given the outliers in the data. Collectively modeling the curves that represent all females (males) in the prefectures can better ease the influence of the few outlying curves than jointly modeling the series for females and males in each prefecture. This finding is because natural disasters are likely to affect females and males in an area equally. Comparing RMSFEs across all sub-national levels, we recommend grouping age-specific mortality rates of the same sex for each region and prefecture, applying the dynamic multivariate functional time series method to obtain base forecasts, and then reconciling the forecasts using the bottom-up method.

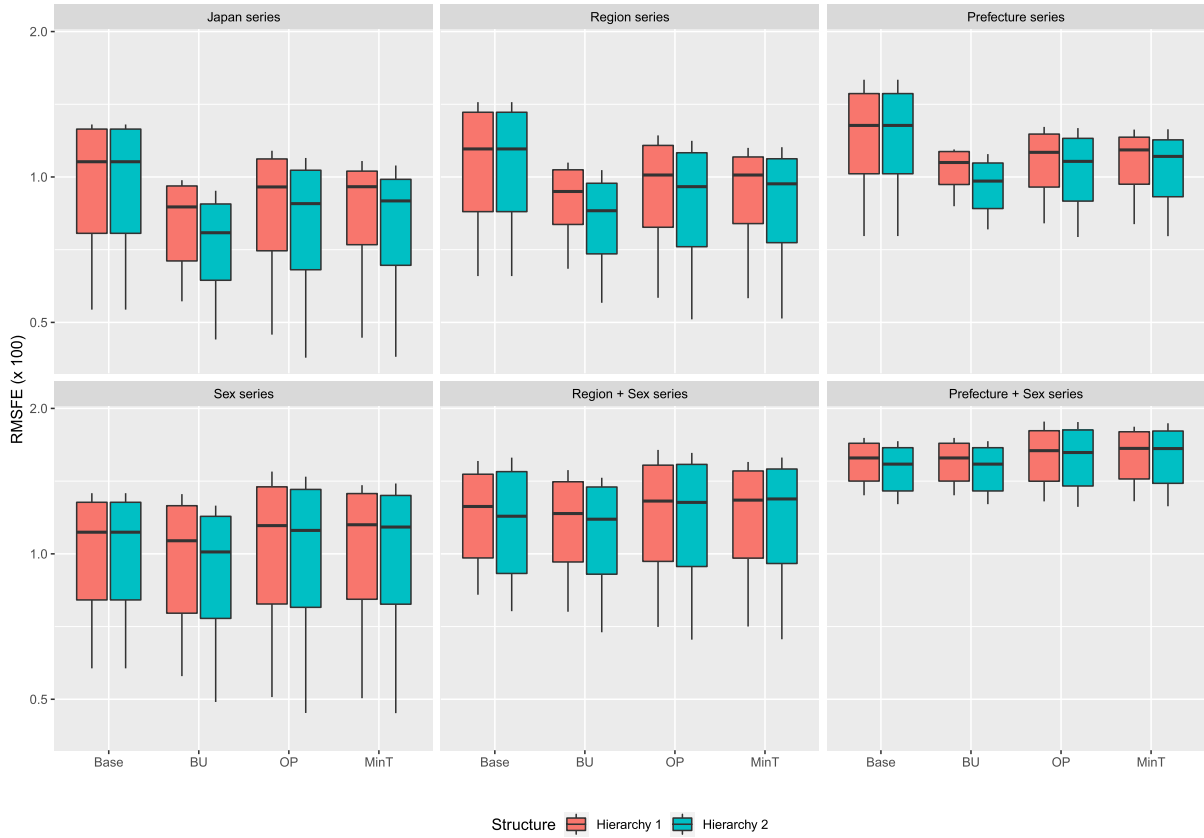


Figure 6: RMSFEs ( $\times 100$ ) for the base and reconciled multivariate functional time series forecasts obtained from different group structures.

### 4.2 Interval Forecast Evaluation and Comparison

In addition to point forecasts, we also evaluate the pointwise interval forecast accuracy using the interval score of Gneiting and Raftery (2007) (see also Gneiting and Katzfuss, 2014). For each year in the forecasting period, the  $h$ -step-ahead prediction intervals are calculated at the  $100(1 - \alpha)\%$  nominal coverage probability. We consider the common case of a symmetric  $100(1 - \alpha)\%$  prediction interval whose lower and upper bounds are predictive quantiles at  $\alpha/2$  and  $1 - \alpha/2$ , denoted by  $\hat{f}_{n+\zeta}^{lb}(x_i)$  and  $\hat{f}_{n+\zeta}^{ub}(x_i)$ , respectively. As defined by Gneiting and Raftery (2007), a scoring rule for the pointwise interval forecast at time point  $x_i$  is

$$\begin{aligned}
 P_\alpha [\hat{f}_{n+\zeta}^{lb}(x_i), \hat{f}_{n+\zeta}^{ub}(x_i); f_{n+\zeta}(x_i)] &= [\hat{f}_{n+\zeta}^{ub}(x_i) - \hat{f}_{n+\zeta}^{lb}(x_i)] \\
 &+ \frac{2}{\alpha} [\hat{f}_{n+\zeta}^{lb}(x_i) - f_{n+\zeta}(x_i)] \mathbb{1} \{f_{n+\zeta}(x_i) < \hat{f}_{n+\zeta}^{lb}(x_i)\} \\
 &+ \frac{2}{\alpha} [f_{n+\zeta}(x_i) - \hat{f}_{n+\zeta}^{ub}(x_i)] \mathbb{1} \{f_{n+\zeta}(x_i) > \hat{f}_{n+\zeta}^{ub}(x_i)\},
 \end{aligned}$$

where  $\alpha$  denotes the level of significance; typically  $\alpha = 0.2$ . It can be easily seen that the optimal interval score is achieved when  $f_{n+\zeta}(x_i)$  lies between  $\hat{f}_{n+\zeta}^{lb}(x_i)$  and  $\hat{f}_{n+\zeta}^{ub}(x_i)$ , with the distance between the upper bound and the lower bound being minimal.

The scoring rule of Gneiting and Raftery (2007) is considered in this study since it combines the coverage rate and halfwidth of the prediction interval. We aim to construct prediction intervals where the empirical coverage rate is close to the nominal one, with a minimum width of the prediction interval. When the empirical coverage rate is less than the nominal one, the constructed prediction interval is narrower than the oracle. This indicates that we are under-estimating the forecast uncertainty. Conversely, if the empirical coverage rate is more than the nominal one, our prediction interval is wider than the oracle. This indicates we are over-estimating the forecast uncertainty, and our prediction intervals are not that informative.

Averaging over different points in a curve and different years in the forecasting period, the mean interval score is defined as

$$\bar{P}_\alpha(h) = \frac{1}{101 \times (16 - h)} \sum_{\zeta=h}^{15} \sum_{i=1}^{101} P_\alpha[\hat{f}_{n+\zeta}^{lb}(x_i), \hat{f}_{n+\zeta}^{ub}(x_i); f_{n+\zeta}(x_i)],$$

where  $P_\alpha[\hat{f}_{n+\zeta}^{lb}(x_i), \hat{f}_{n+\zeta}^{ub}(x_i); f_{n+\zeta}(x_i)]$  denotes the interval score at the  $\zeta^{th}$  curve of the forecasting period. For 15 different forecast horizons, we use a mean statistic given by

$$\text{Mean}(\bar{P}_\alpha) = \frac{1}{15} \sum_{h=1}^{15} \bar{P}_\alpha(h)$$

to evaluate the pointwise interval forecast accuracy.

Figure 7 summarizes the mean interval scores of the univariate and multivariate functional time series applied to the Japanese age-specific mortality rates under Hierarchy 1. Considering multiple mortality curves allows for borrowing information of related populations but at the expense of sacrificing specific information for each sub-population to a certain extent. Hence, the multivariate forecasting method may not always outperform the univariate forecasting method because of the relatively large heterogeneity of the considered mortality series. In our study, pointwise interval forecasts obtained by the dynamic multivariate functional time series method are generally more accurate than those produced by the dynamic univariate functional time series method; the superior forecast accuracy of the dynamic multivariate functional time series method becomes more evident if used in combination with the bottom-up method.

The second and third panels in Figure 7 depict interval forecast accuracies for the total series (i.e., the aggregate mortality of females and males) at region and prefecture levels, respectively. The DMFTS method collectively models all the considered total series to estimate the long-run covariance of the entire population of Japan before producing forecasts. Significant heterogeneity exists in the total series at the Prefecture or Region levels since earthquakes and tsunamis influenced mortality at several prefectures during the considered period. Hence, natural disaster-related deaths as outliers tend to inflate the estimation variance of every series used in the DMFTS modeling process. In contrast, the base method considers one total series, either for a region or a prefecture, at a time. Only those prefectures or regions with outlying death counts would see additional estimation variance. Moreover, the dynamic modeling method may not accurately recover the long-run covariance of mortality series of all Japanese residents based on a finite sample. Due to these reasons, the DMFTS method sometimes performs inferiorly to the independent method. Applying any of the considered reconciliation methods improves interval forecast accuracies of the total series at region and prefecture levels, as shown in Figure 7. The forecast accuracy improvement is because Japan's population hierarchy helps stabilize the abnormal forecasts inflated by excess uncertainty in estimation.



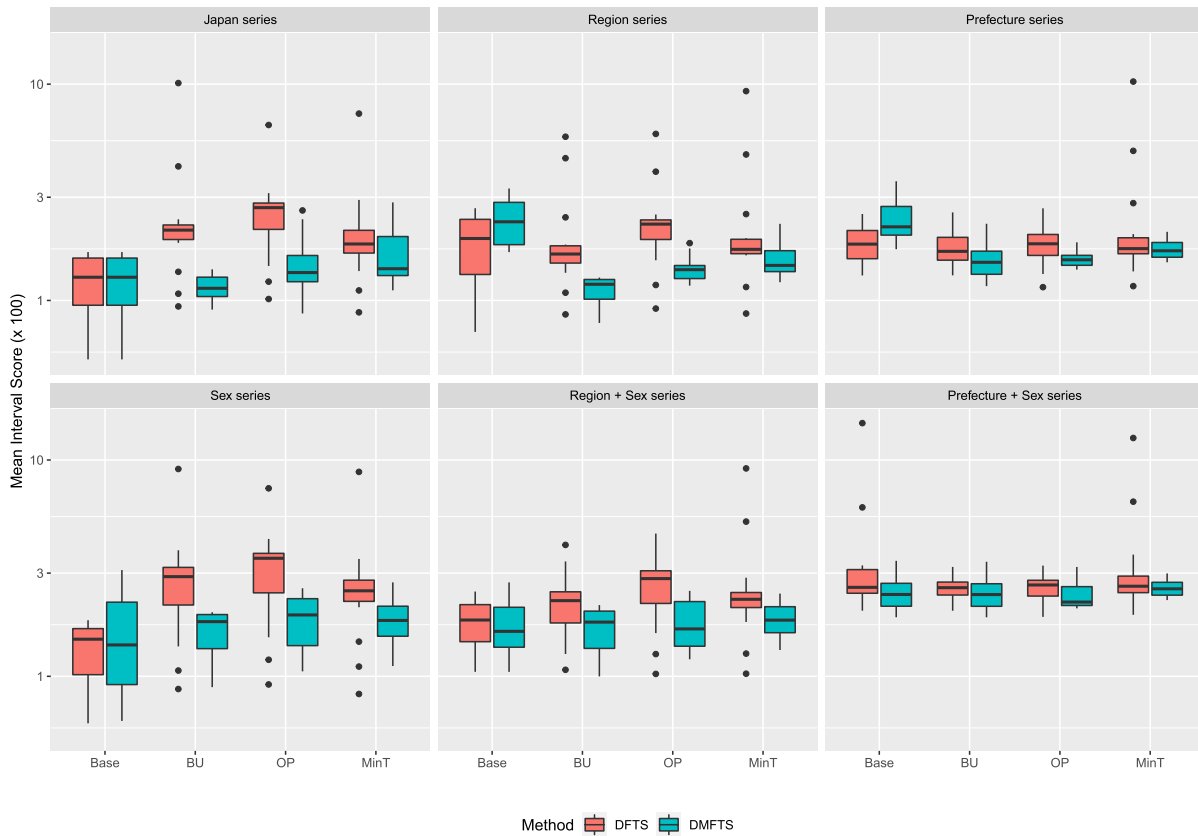


Figure 7: Interval forecasting performance of the base and reconciled forecasts for the univariate and multivariate functional time series methods.

Figure 8 compares the mean interval scores for the base and reconciled multivariate functional time series forecasts obtained using two disaggregation hierarchies. The result of Hierarchy 1 dominating Hierarchy 2, shown in the figure, can be explained as follows. By applying the dynamic multivariate functional time series method under the Hierarchy 2 structure, it focuses on common features of female (male) mortality, leading to inaccurate pointwise prediction intervals for outlying series due to losing specific information for the sub-population. Given that interval scores will be inflated by every holdout value outside the calculated prediction interval, outlier curves cause a greater penalty to Hierarchy 2 than to Hierarchy 1. Hence, jointly modeling the mortality series for females (males) of each prefecture and region is better for improving pointwise interval forecast accuracy.

## 5 Conclusion

When applied to the functional time series formed by disaggregated series, such as sub-national age-specific mortality rates, the dynamic multivariate functional time series method can be combined with reconciliation methods to obtain improved point and interval forecasts.

The dynamic multivariate functional time series method is applied to data on Japanese age-specific mortality rates from 1975 to 2016. The dynamic multivariate functional time series

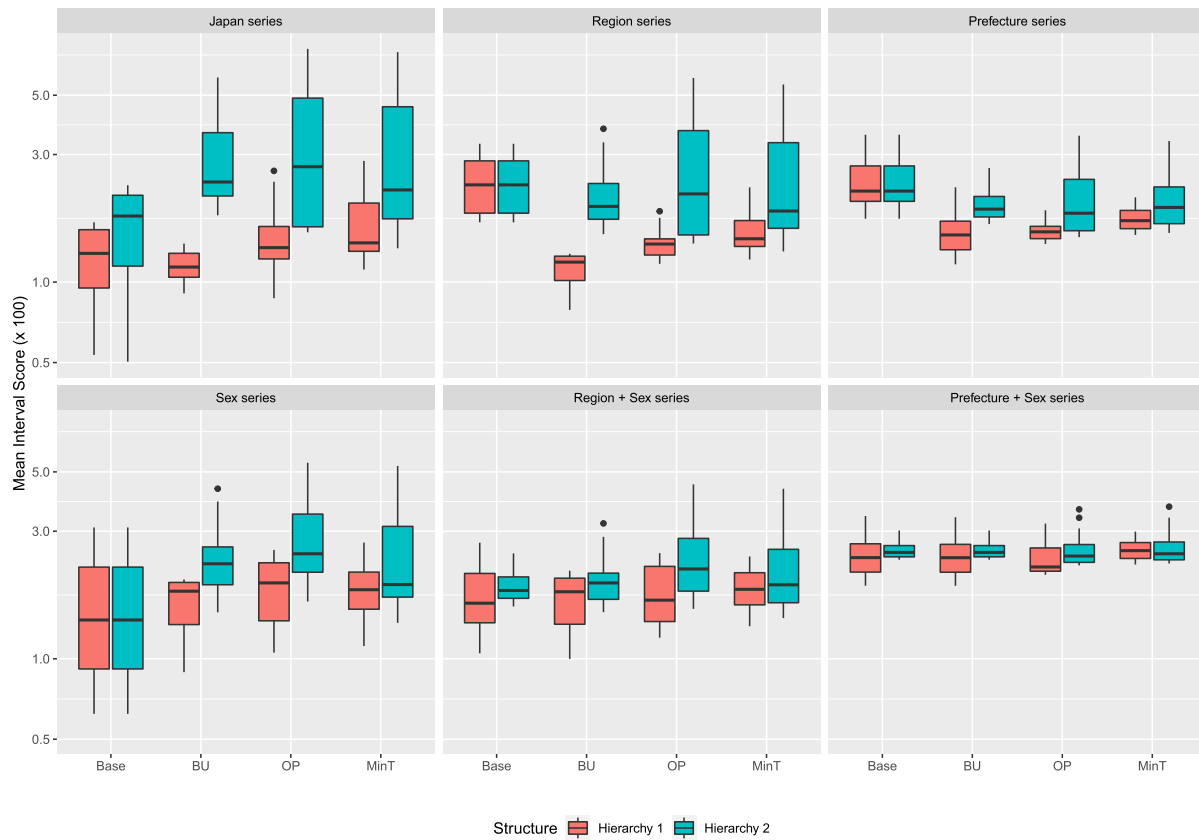


Figure 8: Mean interval scores for the base and reconciled multivariate functional time series forecasts obtained from different group structures.

method produces the most accurate point and interval forecasts of Japanese mortality rates combined with the bottom-up reconciliation method. Different group structures involving geographical factors and sex are considered and compared in forecasting and reconciliation. For point forecasting, we recommend that, first, the national mortality data are disaggregated by sex, and further, each series is split according to a geographical area. Because of the bias-variance trade-off, reversing the disaggregation order, that is, first, according to the geographical area, and then, by sex, yields more accurate pointwise interval forecasts, evaluated by the mean interval scores.

There are some ways in which this paper can be further extended. Here, we briefly mention three:

- 1) We have considered two factors, namely sex and geographical locations, to disaggregate age-specific mortality rates. It is possible to enrich the information contained in the group structure by differentiating these rates according to other factors, such as the cause of death (Murray and Lopez, 1997; Gaille and Sherris, 2015), social and economic aspects (Bassuk et al., 2002; Singh et al., 2013) and native-born or immigrant status (Singh and Siahpush, 2001). If appropriate data are available, we may attempt to extend the proposed grouped multivariate functional time series forecasting method to cause-specific mortality rates and occupation-specific mortality rates.
- 2) We expect the mortality functions at the prefecture-level to be homogeneous so that the

empirical long-run covariance function can be accurately estimated. When there are major natural disasters such as earthquakes and tsunamis in some prefectures, the observed mortality functions may have different shapes. Such functions of strange shapes can be viewed as outliers. Functional depth-based outlier detection methods of Hyndman and Shang (2010), and the recently developed novel multivariate functional outlier detection methods of Dai and Genton (2018) and Dai et al. (2020) can be adopted to check if a particular curve is an outlying observation. Among other methods, the robust functional principal component analysis method and the robust regularized singular value decomposition method of Shang et al. (2019) can be applied to mitigate influences of functional outliers in modeling and forecasting age-specific mortality rates (see also Zhang et al., 2013; Bali et al., 2011).

- 3) We evaluate and compare the pointwise interval forecast accuracy. To our best knowledge, uniform prediction intervals are still rarely used in the context of functional time series models, owing to the difficulty in computing an appropriate semi-metric distance. A future extension can consider the uniform prediction intervals of grouped functional time series and study the interval forecast accuracy in function space.

## Supplementary Material

For reproducibility, all R codes are collated in a Github repository at <https://github.com/hanshang/GMFTS>. This repository contains the following files:

- **README:** Describe the functionality of each R file.
- **R files with indices 1 to 29:** Functions used to compute numerical results presented in the paper.
- **Example:** Contains a quick example of forecasting prefecture-level age-specific mortality rates using the DMFTS method.

## Acknowledgement

The authors are very grateful to the editor and two reviewers for their constructive and insightful comments that lead to an improvement of the paper.

## References

- Aue A, Norinho DD, Hörmann S (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association: Theory and Methods*, 110(509): 378–392.
- Bali JL, Boente G, Tyler DE, Wang J-L, et al. (2011). Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, 39(6): 2852–2882.
- Bassuk SS, Berkman LF, Amick III BC (2002). Socioeconomic status and mortality among the elderly: Findings from four US communities. *American Journal of Epidemiology*, 155(6): 520–533.
- Booth H, Tickle L (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, 3(1–2): 3–43.
- Cairns AJ, Blake D, Dowd K, Coughlan GD, Khalaf-Allah M M (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA*, 41(1): 29–59.

- Chatfield C (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2): 121–135.
- Chiou J-M (2012). Dynamical functional prediction and classification with application to traffic flow prediction. *Annals of Applied Statistics*, 6(4): 1588–1614.
- Chiou J-M, Chen Y-T, Yang Y-F (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 24(4): 1571–1596.
- Coulmas F (2007). *Population Decline and Ageing in Japan – The Social Consequences*. Routledge, New York.
- Currie ID, Durban PHC, Eilers M (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4): 279–298.
- Dai W, Genton MG (2018). Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4): 923–934.
- Dai W, Mrkvička T, Sun Y, Genton MG (2020). Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, 149: 106960.
- Dangerfield BJ, Morris JS (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting*, 8(2): 233–241.
- Dowd K, Cairns AJ, Blake D, Coughlan GD, Khalaf-Allah M (2011). A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2): 334–356.
- Gaille SA, Sherris M (2015). Causes-of-death mortality: What do we know on their dependence? *North American Actuarial Journal*, 19(2): 116–128.
- Gao Y, Shang HL, Yang Y (2019). High-dimensional functional time series forecasting: An application to age-specific mortality rates. *Journal of Multivariate Analysis*, 170: 232–243.
- Gneiting T, Katzfuss M (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1: 125–151.
- Gneiting T, Raftery AE (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102(477): 359–378.
- Hall P, Vial C (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 68(4): 689–705.
- Happ C, Greven S (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association: Theory and Methods*, 113(522): 649–659.
- Hatzopoulos P, Haberman S (2013). Common mortality modeling and coherent forecasts. an empirical analysis of worldwide mortality data. *Insurance, Mathematics & Economics*, 52(2): 320–337.
- Hyndman RJ, Ahmed RA, Athanasopoulos G, Shang HL (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9): 2579–2589.
- Hyndman RJ, Shang HL (2009). Forecasting functional time series (with discussions). *Journal of the Korean Statistical Society*, 38(3): 199–221.
- Hyndman RJ, Shang HL (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1): 29–45.
- Hyndman RJ, Ullah MS (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10): 4942–4956.
- Jacques J, Preda C (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71: 92–106.
- Japanese Mortality Database (2017). *National Institute of Population and Social Security Research*. Available at <http://www.ipss.go.jp/p-toukei/JMD/index-en.html> (data downloaded

- on July 18, 2016).
- Jarner SF, Kryger EM (2011). Modelling adult mortality in small populations: The saint model. *ASTIN Bulletin: The Journal of the IAA*, 41(2): 377–418.
- Kolkiewicz A, Rice G, Xie Y (2021). Projection pursuit based tests of normality with functional data. *Journal of Statistical Planning and Inference*, 211: 326–339.
- Murray CJL, Lopez AD (1997). Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. *The Lancet*, 349(9064): 1498–1504.
- OECD (2020). Pensions at a glance 2019: OECD and G20 indicators, Technical report, OECD Publishing. Retrieved from [http://https://www.oecd-ilibrary.org/social-issues-migration-health/pensions-at-a-glance-2019\\_b6d3dcfc-en](http://https://www.oecd-ilibrary.org/social-issues-migration-health/pensions-at-a-glance-2019_b6d3dcfc-en) on November 4, 2020.
- Perla F, Richman R, Scognamiglio S, Wüthrich MV (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7): 572–598.
- Rice G, Shang HL (2017). A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. *Journal of Time Series Analysis*, 38(4): 591–609.
- Rice J, Silverman B (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 53(1): 233–243.
- Russolillo M, Giordano G, Haberman S (2011). Extending the Lee–Carter model: A three-way decomposition. *Scandinavian Actuarial Journal*, 2011(2): 96–117.
- Sefton J, Weale M (2009). *Reconciliation of National Income and Expenditure: Balanced Estimates of National Income for the United Kingdom 1920–1990*. Cambridge University Press, Cambridge.
- Shang HL (2016). Mortality and life expectancy forecasting for a group of populations in developed countries: A multilevel functional data method. *Annals of Applied Statistics*, 10(3): 1639–1672.
- Shang HL (2017). Reconciling forecasts of infant mortality rates at national and sub-national levels: Grouped time-series methods. *Population Research and Policy Review*, 36(1): 55–84.
- Shang HL (2018). Bootstrap methods for stationary functional time series. *Statistics and Computing*, 28(1): 1–10.
- Shang HL (2019). Dynamic principal component regression: Application to age-specific mortality forecasting. *ASTIN Bulletin: The Journal of the IAA*, 49(3): 619–645.
- Shang HL, Hyndman RJ (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, 26(2): 330–343.
- Shang HL, Kearney F (2021). Dynamic functional time-series forecasts of foreign exchange implied volatility surfaces. *International Journal of Forecasting*. In press.
- Shang HL, Yang Y, Kearney F (2019). Intraday forecasts of a volatility index: Functional time series methods with dynamic updating. *Annals of Operations Research*, 282(1): 331–354.
- Singh GK, Azuine RE, Siahpush M, Kogan MD (2013). All-cause and cause-specific mortality among US youth: Socioeconomic and rural-urban disparities and international patterns. *Journal of Urban Health*, 90(3): 388–405.
- Singh GK, Siahpush M (2001). All-cause and cause-specific mortality of immigrants and native born in the United States. *American Journal of Public Health*, 91(3): 392.
- Stone R, Champenowne DG, Meade JE (1942). The precision of national income estimates. *The Review of Economic Studies*, 9(2): 111–125.

- The Government of Japan (2021). *Japan's Regional Strength*. Available at <https://www.japan.go.jp/regions/index.html> (Accessed August 26, 2021).
- Villegas AM, Haberman S, Kaishev VK, Millossovich P (2017). A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin: The Journal of the IAA*, 47(3): 631–679.
- Wickramasuriya SL, Athanasopoulos G, Hyndman RJ (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association: Theory and Methods*, 114(526): 804–819.
- Wood SN (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 65(1): 95–114.
- Yao F, Müller H-G, Wang J-L (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association: Theory and Methods*, 100(470): 577–590.
- Zellner A, Tobias J (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, 19(5): 457–465.
- Zhang L, Shen H, Huang JZ (2013). Robust regularized singular value decomposition with application to mortality data. *Annals of Applied Statistics*, 7(3): 1540–1561.
- Zivot E, Wang J (2006). *Modeling Financial Time Series with S-PLUS*. Springer, New York.