MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Grouped functional time series forecasting: an application to age-specific mortality rates

Han Lin Shang and Rob J Hyndman

February 2016

Working Paper 04/16

Grouped functional time series forecasting: an application to age-specific mortality rates

Han Lin Shang
Research School of Finance, Actuarial Studies & Statistics
Australian National University
Canberra ACT 2601
Australia
Email: hanlin.shang@anu.edu

Rob J Hyndman
Department of Econometrics & Business Statistics
Monash University
Clayton VIC 3800
Australia
Email: rob.hyndman@monash.edu

29 February 2016

JEL classification: C14, C32, J11

Grouped functional time series forecasting: an application to age-specific mortality rates

Abstract

Age-specific mortality rates are often disaggregated by different attributes, such as sex, state and ethnicity. Forecasting age-specific mortality rates at the national and sub-national levels plays an important role in developing social policy. However, independent forecasts of age-specific mortality rates at the sub-national levels may not add up to the forecasts at the national level. To address this issue, we consider the problem of reconciling age-specific mortality rate forecasts from the viewpoint of grouped univariate time series forecasting methods (Hyndman, Ahmed, et al., 2011), and extend these methods to functional time series forecasting, where age is considered as a continuum. The grouped functional time series methods are used to produce point forecasts of mortality rates that are aggregated appropriately across different disaggregation factors. For evaluating forecast uncertainty, we propose a bootstrap method for reconciling interval forecasts. Using the regional age-specific mortality rates in Japan, obtained from the Japanese Mortality Database, we investigate the one- to ten-step-ahead point and interval forecast accuracies between the independent and grouped functional time series forecasting methods. The proposed methods are shown to be useful for reconciling forecasts of age-specific mortality rates at the national and sub-national levels, and they also enjoy improved forecast accuracy averaged over different disaggregation factors.

Keywords: forecast reconciliation; hierarchical time series forecasting; bottom-up; optimal combination; Japanese Mortality Database

1 Introduction

Functional time series often consist of random functions observed at regular time intervals. Depending on whether or not the continuum is also a time variable, functional time series can be grouped into two categories. On one hand, functional time series can arise by separating an almost continuous time record into natural consecutive intervals such as days, months or years (see Hörmann and Kokoszka, 2012). Examples include daily price curves of a financial stock (Kokoszka and Zhang, 2012), and monthly sea surface temperature in climatology (Shang and Hyndman, 2011). On the other hand, functional time series can also arise when observations in a time period can be considered together as finite realizations of an underlying continuous function; for example, annual age-specific mortality rates in demography (e.g., Chiou and Müller, 2009; Hyndman and Ullah, 2007).

In either case, the functions obtained form a time series $\{\mathcal{X}_t,\ t \in Z\}$, where each $\mathcal{X}_t$ is a (random) function $\mathcal{X}_t(z)$ and $z \in \mathcal{I}$ represents a continuum on a finite interval. We refer to such data structures as functional time series.

There has been a rapidly growing body of research on functional time series forecasting methods. From a parametric viewpoint, Bosq, 2000 proposed the functional autoregressive process of order 1 and derived one-step-ahead forecasts that are based on a regularized form of the Yule-Walker equations. From a nonparametric perspective, Besse, Cardot, and Stephenson, 2000 proposed functional kernel regression to measure the temporal dependence via a similarity measure characterized by neighborhood distance (also known as semi-metric), kernel function and bandwidth. Hyndman and Ullah, 2007 use functional principal component analysis to decompose smoothed functional time series into a set of functional principal components and their associated principal component scores. The temporal dependency in the original functional time series is inherited by the correlation within each principal component score and the possible cross-correlations between principal component scores. Hyndman and Ullah, 2007 applied univariate time series forecasting models to forecast these scores individually, while Aue, Norinho, and Hörmann, 2015 considered a multivariate time series forecasting method to capture any correlations between principal component scores. Both univariate and multivariate time series forecasting methods have their own advantages and disadvantages (see Aue, Norinho, and Hörmann, 2015; Peña and Sánchez, 2007, for a comparison).

In this paper, we also use functional principal component regression as a forecasting technique, applied to a large multivariate set of functional time series with rich structure. There have been relatively few research contributions dealing with multivariate functional time series forecasting (see for example, Chiou, Yang, and Chen, 2015; Kowal, Matteson, and Ruppert, 2015). To our knowledge, there has been no study that takes account of aggregation constraints within multivariate functional time series forecasting. This is the gap we wish to address.

To be specific, we consider age-specific mortality rates observed annually as an example of a functional time series, where the continuum is the age variable. These age-specific mortality rates can be observed at the national level, and can be disaggregated by various attributes such as sex, state or ethnicity. Forecasts are often required for national mortality, as well as sub-national mortality disaggregated by different attributes. When a functional forecasting method is applied to each subset, the sum of the forecasts will not generally add up to the forecasts obtained by applying the method to the aggregated national data.

This problem is known as forecast reconciliation, which has been addressed for univariate time series forecasting. Sefton and Weale, 2009 considered forecast reconciliation in the context of national account balancing, while Hyndman, Ahmed, et al., 2011 demonstrated the usefulness of forecast reconciliation methods in the context of tourist demand. In this paper, we develop reconciliation methods tailored for multivariate functional time series.

We put forward two statistical methods, namely bottom-up and optimal combination methods, to reconcile point and interval forecasts of age-specific mortality, and potentially improve the point and interval forecast accuracies. The bottom-up method involves forecasting each of the disaggregated series and then using simple aggregation to obtain forecasts for the aggregated series (Kahn, 1998). This method works well where the bottom-level series have high signal-to-noise ratio. For highly disaggregated series, this does not tend to work well as the series become too noisy; also, any relationships between series are ignored. This motivated the development of an optimal combination method (Hyndman, Ahmed, et al., 2011), where forecasts are obtained independently for all series at all levels of disaggregation and then a linear regression model is used with a generalized least-squares estimator to optimally combine and reconcile these forecasts. We propose a modification of this approach for use with functional time series.

Using the national and sub-national Japanese age-specific mortality rates from 1975 to 2013, we compare the point and interval forecast accuracies among the independent forecasting, bottom-up and optimal combination methods. For evaluating the point forecast accuracy, we consider the mean absolute forecast and root mean squared forecast errors, and found that the bottom-up method gives the most accurate overall point forecasts. For evaluating the interval forecast accuracy, we use the mean interval score, and again found that the bottom-up method gives the most accurate overall interval forecasts.

The rest of this paper is structured as follows. In Section 2, we describe the motivating data set, which is Japanese national and sub-national age-specific mortality rates. In Section 3, we describe the functional principal component regression for producing point and interval forecasts, then introduce grouped functional time series forecasting methods in Section 4. We evaluate and compare point and interval forecast accuracies between the independent and grouped functional time series forecasting methods in Sections 5 and 6. Conclusions are presented in Section 7, along with some reflections on how the methods presented here can be further extended.

2 Japanese age-specific mortality rates for 47 prefectures

In many developed countries such as Japan, increases in longevity and an aging population have led to concerns regarding the sustainability of pensions, health and aged care systems (see, for example, Coulmas, 2007; OECD, 2013). These concerns have resulted in a surge of interest amongst government policy makers and planners in accurately modeling and forecasting age-specific mortality rates. Sub-national forecasts of age-specific mortality rates are useful for informing policy within local regions. Any improvement in the forecast accuracy of mortality rates will be beneficial for determining the allocation of current and future resources at the national and sub-national levels.

We study Japanese age-specific mortality rates from 1975 to 2013, obtained from the Japanese Mortality Database (2015). We consider ages from 0 to 99 in single years of age, while the last age group contains all ages at and beyond 100. The structure of the data is displayed in Table 1 where each row denotes a level of disaggregation. At the top level, we have total age-specific mortality rates for Japan. We can split these total mortality rates by sex, by region, or by prefecture. There are eight regions in Japan, which contain a total of 47 prefectures. The most disaggregated data

arise when we consider the mortality rates for each combination of prefecture and sex, giving a total of $47 \times 2 = 94$ series. In total, across all levels of disaggregation, there are 168 series.

Table 1: *Hierarchy of Japanese mortality rates.*

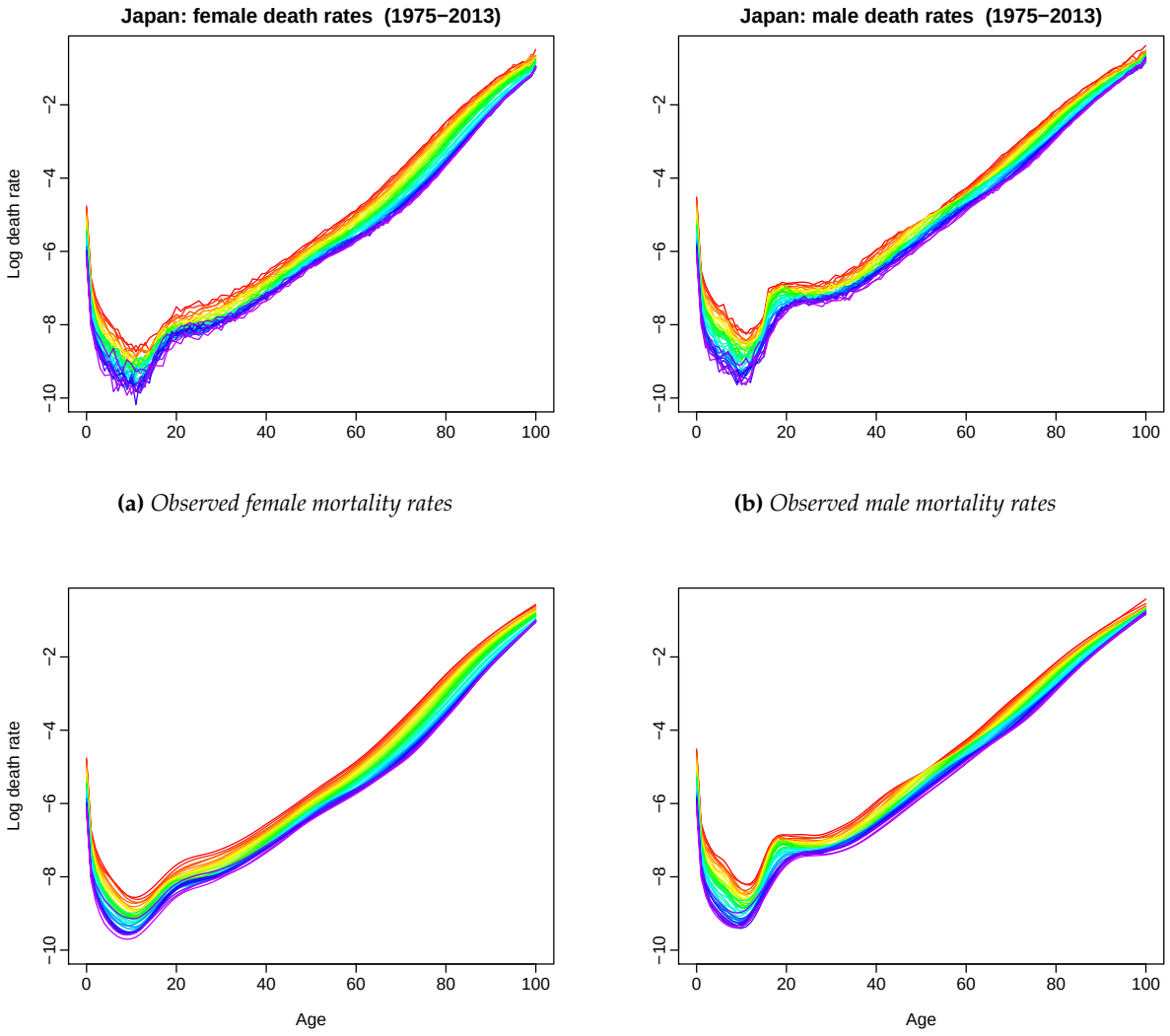
Level	Number of series
Japan	1
Sex	2
Region	8
Sex $\times$ Region	16
Prefecture	47
Sex $\times$ Prefecture	94
Total	168

2.1 Rainbow plots

Figure 1 shows rainbow plots of the female and male age-specific log mortality rates in Japan from 1975 to 2013 (Hyndman and Shang, 2010). The time ordering of the curves follows the color order of a rainbow, where curves from the distant past are shown in red and the more recent curves are shown in purple. The figures show typical mortality curves for a developed country, with rapidly decreasing mortality rates in the early years of life, followed by an increase during the teenage years, a plateau for young adults, and then a steady increase from about the age of 30. Females have lower mortality rates than males at all ages.

From Figures 1a and 1b, the observed mortality rates are not smooth across age due to observational noise. To obtain smooth functions and deal with possible missing values, we consider a penalized regression spline smoothing with monotonic constraint, described in Section 3.2. It takes into account the shape of log mortality curves (see also D'Amato, Piscopo, and Russolillo, 2011; Hyndman and Ullah, 2007).

Figures 1c and 1d demonstrate smooth age-specific mortality rates for Japan females and males, and we apply smoothing to all series at different levels of disaggregation. We have developed a Shiny app (Chang et al., 2015) in R (R Core Team, 2015) to allow interactive exploration of the smoothing of all the data; this is available in the online supplementary material.



(a) *Observed female mortality rates*

(b) *Observed male mortality rates*

(c) *Smoothed female mortality rates*

(d) *Smoothed male mortality rates*

Figure 1: *Functional time series graphical displays*

2.2 Image plots

Another visual perspective of the data is shown in the image plots of Figure 2. Here we graph the log of the ratio of mortality rates for each prefecture to mortality rates for the whole country, thus allowing relative mortality comparisons to be made. A divergent color palette is used with blue representing positive values and orange denoting negative values. The prefectures are ordered geographically from north to south, so the most northerly prefecture (Hokkaidō) is at the top of the panels, and the most southerly prefecture (Okinawa) is at the bottom of the panels.

The top row of panels shows mortality rates for each prefecture and age, averaged over all years. Several striking features become apparent. There are strong differences between the

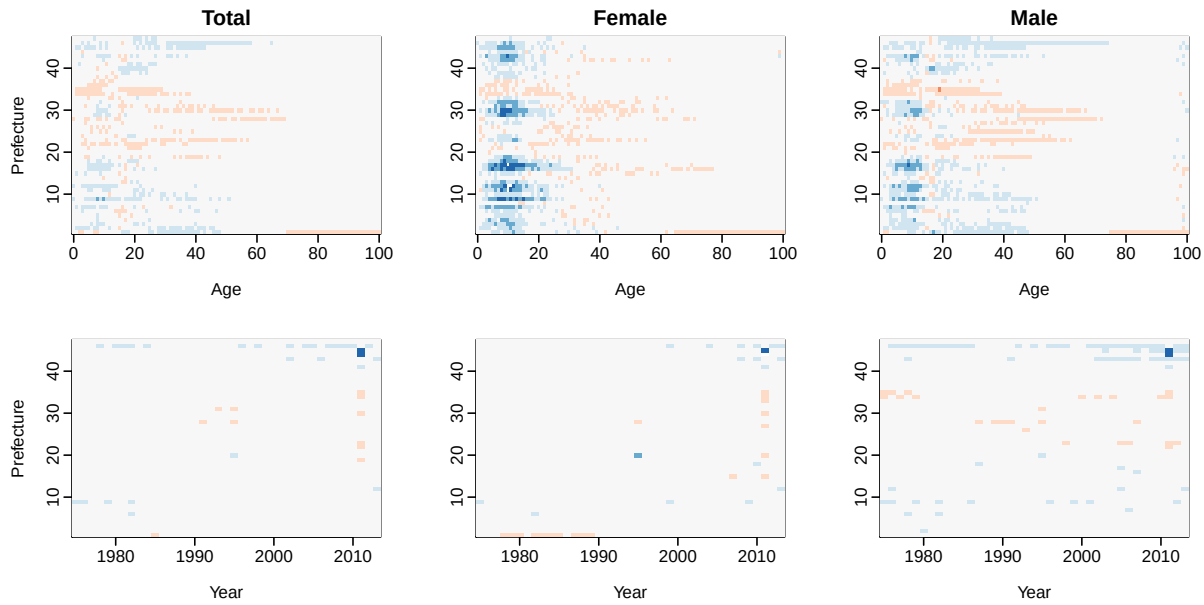


Figure 2: *Image plots showing log ratios of mortality rates. The top panel shows mortality rates averaged over years, while the bottom panel shows mortality rates averaged over ages. Prefectures are numbered geographically from north to south.*

prefectures for children, especially females; this is possibly due to socio-economic differences, and accessibility of health services. The most southerly prefecture of Okinawa has particularly low mortality rates for older people; this is consistent with the extreme longevity for which Okinawa is famous (see for example, Suzuki, Willcox, and Willcox, 2004; Takata et al., 1987; Willcox et al., 2007).

The bottom row of panels shows mortality rates for each prefecture and year, averaged over all ages. Here there is less information to be seen, but three outliers are highlighted. In 2011, in prefectures 44 (Miyagi) and 45 (Iwate) there was a large increase in mortality compared to other prefectures. These are northern coastal regions, and the inflated relative mortality is due to the tsunami of 11 March 2011. There is a corresponding decrease in relative mortality in some other provinces.

In 1995, there is an increase in mortality for prefecture 20 (Hyōgo). This corresponds with the Kobe (Great Hanshin) earthquake of 17 January 1995.

Also evident is the decreased female mortality in Okinawa up until 1990, perhaps suggesting a recent decline in the relative mortality advantages enjoyed by residents in this region.

3 Methodology

3.1 Functional principal component analysis

Let $(\mathcal{X}_t : t \in Z)$ be an arbitrary functional time series. It is assumed that the observations $\mathcal{X}_t$ are elements of the Hilbert space $\mathcal{H} = L^2(\mathcal{I})$ equipped with the inner product $\langle w, v\rangle = \int_{\mathcal{I}} w(z)v(z)dz$, where z represents a continuum and $\mathcal{I}$ represents the function support range. Each function is a square integrable function satisfying $\|\mathcal{X}_t\|^2 = \int_{\mathcal{I}} \mathcal{X}_t^2(z)dz < \infty$ and associated norm. All random functions are defined on a common probability space (Ω, A, P). The notation $\mathcal{X} \in L^p_{\mathcal{H}}(\Omega, A, P)$ is used to indicate $\mathrm{E}(\|\mathcal{X}\|^p) < \infty$ for some $p > 0$. When $p = 1$, $\mathcal{X}(z)$ has the mean curve $\mu(z) = \mathrm{E}[\mathcal{X}(z)]$; when $p = 2$, a non-negative definite covariance function is given by

$$c_{\mathcal{X}}(y,z) = \mathrm{Cov}[\mathcal{X}(y), \mathcal{X}(z)] = \mathrm{E}\{[\mathcal{X}(y) - \mu(y)][\mathcal{X}(z) - \mu(z)]\} \tag{1}$$

for all $y, z \in \mathcal{I}$. The covariance function $c_{\mathcal{X}}(y,z)$ in (1) allows the covariance operator of $\mathcal{X}$, denoted by $\mathcal{K}_{\mathcal{X}}$ to be defined as

$$\mathcal{K}_{\mathcal{X}}(\phi)(z) = \int_{\mathcal{I}} c_{\mathcal{X}}(y,z)\phi(y)dy.$$

Via Mercer's lemma, there exists an orthonormal sequence (ϕ_k) of continuous function in $L^2(\mathcal{I})$ and a non-increasing sequence λ_k of positive numbers such that

$$c_{\mathcal{X}}(y,z) = \sum_{k=1}^{\infty} \lambda_k \phi_k(y)\phi_k(z), \qquad y, z \in \mathcal{I}.$$

By the separability of Hilbert spaces, the Karhunen–Loève expansion of a stochastic process $\mathcal{X}(z)$ can be expressed as

$$\begin{aligned}\mathcal{X}(z) &= \mu(z) + \sum_{k=1}^{\infty} \beta_k \phi_k(z) \\ &= \mu(z) + \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k \phi_k(z),\end{aligned}$$

where $\xi_k = 1/\sqrt{\lambda_k} \int_{\mathcal{I}} [\mathcal{X}(z) - \mu(z)]\phi_k(z)dz$ is an uncorrelated random variable with zero mean and unit variance. The principal component scores $\beta_k = \sqrt{\lambda_k}\xi_k$ are given by the projection of $\mathcal{X}(z) - \mu(z)$ in the direction of the k^{th} eigenfunction ϕ_k, i.e., $\beta_k = \langle \mathcal{X}(z) - \mu(z), \phi_k(z)\rangle$.

As a widely used dimension reduction technique, functional principal component analysis summarizes the main features of an infinite-dimensional object by a few basis functions. For theoretical, methodological, and applied aspects of functional principal component analysis, consult the survey articles by Hall, 2011, Shang, 2014, Wang, Chiou, and Müller, 2015 and Reiss et al., 2016.

3.2 Nonparametric smoothing technique

Functional data are intrinsically infinite dimensional, although we can only observe functional data at dense grid points (see for example, Ramsay and Silverman, 2005) or sparse grid points (see for example, Müller, 2005). In practice, the observed data are often contaminated by random noise, referred to as measurement errors. As defined by Wang, Chiou, and Müller, 2015, measurement errors can be viewed as random fluctuations around a continuous and smooth function, or as actual errors in the measurement.

We assume that there are underlying L_2 continuous and smooth functions $\mathcal{X}_t(z)$ such that

$$\mathcal{Y}_t(z_j) = \mathcal{X}_t(z_j) + \sigma_t(z_j)\varepsilon_{t,j}, \qquad t = 1,\dots,n,\ j = 1,\dots,p,$$

where $\mathcal{Y}_t(z_j)$ denotes the raw log mortality rates, $\{\varepsilon_{t,j}\}$ are independent and identically distributed (iid) random variables across t and j with zero mean and unit variance, and $\sigma_t(z_j)$ allows for heteroskedasticity. We observe that measurement errors are realized only at those time points z_j where measurements are being taken. As a result, these errors are treated as discretized data $\varepsilon_{t,j}$. However, in order to estimate the variance $\sigma_t^2(z_j)$, we assume that there is a latent smooth function $\sigma^2(z)$ observed at discrete time points.

Let $m_t(z_j) = \exp[\mathcal{Y}_t(z_j)]$ be the observed central mortality rates for age z_j in year t and define $E_t(z_j)$ to be the population of age z_j at 30 June in year t (often known as the "exposure-at-risk"). The observed mortality rate follows a Poisson distribution with estimated variance

$$\widehat{\sigma}_t^2(z_j) = \frac{1}{m_t(z_j)E_t(z_j)}.$$

For modeling age-specific log mortality, Hyndman and Ullah, 2007 advocated the application of weighted penalized regression splines with a monotonic constraint for ages above 65, where the

weights are equal to the inverse variances, $w_t(z_j) = 1/\widehat{\sigma}_t^2(z_j)$. For each year t,

$$\widehat{\mathcal{X}}_t(z_j) = \underset{\theta_t(z_j)}{\operatorname{argmin}} \sum_{j=1}^{M} w_t(z_j) \left| \mathcal{Y}_t(z_j) - \theta_t(z_j) \right| + \lambda \sum_{j=1}^{M-1} \left| \theta_t'(z_{j+1}) - \theta_t'(z_j) \right|,$$

where z_j represents different ages (grid points) in a total of M grid points, λ represents a smoothing parameter, θ' denotes the first derivative of smooth function θ, which can be both approximated by a set of B-splines (see for example, de Boor, 2001). The L_1 loss function and L_1 penalty function are used to obtain robust estimates. This monotonic constraint helps to reduce the noise from estimation of high ages (see also D'Amato, Piscopo, and Russolillo, 2011; Fang and Härdle, 2015).

3.3 Functional principal component regression

By using functional principal component analysis, a time series of smoothed functions $\boldsymbol{\mathcal{X}}(z) = \{\mathcal{X}_1(z), \dots, \mathcal{X}_n(z)\}$ is decomposed into orthogonal functional principal components and their associated principal component scores, given by

$$\begin{aligned} \mathcal{X}_t(z) &= \mu(z) + \sum_{k=1}^{\infty} \beta_{t,k} \phi_k(z) \\ &= \mu(z) + \sum_{k=1}^{K} \beta_{t,k} \phi_k(z) + e_t(z), \end{aligned} \tag{2}$$

where $\mu(z)$ is the mean function; $\{\phi_1(z), \dots, \phi_K(z)\}$ is a set of the first K functional principal components; $\boldsymbol{\beta}_1 = (\beta_{1,1}, \dots, \beta_{1,n})^\top$ and $\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}$ denotes a set of principal component scores and $\boldsymbol{\beta}_k \sim N(0, \lambda_k)$ where λ_k is the k^{th} eigenvalue of the covariance function in (1); $e_t(z)$ denotes the model truncation error function with mean zero and finite variance; and $K < n$ is the number of retained 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 $\boldsymbol{\mathcal{X}}(z)$ can be adequately summarized by the K-dimensional vector $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$.

Although it can be a research topic on its own, there are several approaches for selecting K: (1) scree plot or the fraction of variance explained by the first few functional principal components (Chiou, 2012); (2) pseudo-versions of Akaike information criterion and Bayesian information criterion (Yao, Müller, and Wang, 2005); (3) predictive cross validation leaving out one or more curves (Rice and Silverman, 1991); (4) bootstrap methods (Hall and Vial, 2006).

Here, the value of K is chosen as the minimum that reaches a certain level of the proportion of total variance explained by the K leading components such that

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

where $\delta = 90\%$, $\mathbb{1}_{\{\widehat{\lambda}_k>0\}}$ is to exclude possible zero eigenvalues, and $\mathbb{1}\{\cdot\}$ represents the binary indicator function.

In a dense and regularly spaced functional time series, the mean function $\widehat{\mu}(z) = \frac{1}{n}\sum_{t=1}^{n} \mathcal{X}_t(z)$ and covariance function $\widehat{c}_{\mathcal{X}}(y,z)$ can be empirically estimated and they are shown to be consistent under the weak dependency (Hörmann and Kokoszka, 2010). From the empirical covariance function, we can extract empirical functional principal component functions $\boldsymbol{\mathcal{B}} = \{\widehat{\phi}_1(z),\dots,\widehat{\phi}_K(z)\}$ using singular value decomposition. Conditioning on the smoothed functions $\boldsymbol{\mathcal{X}}(z) = \{\mathcal{X}_1(z),\dots,\mathcal{X}_n(z)\}$ and the estimated functional principal components $\boldsymbol{\mathcal{B}}$, the h-step-ahead point forecast of $\mathcal{X}_{n+h}(z)$ can be obtained as

$$\widehat{\mathcal{X}}_{n+h|n}(z) = \mathrm{E}[\mathcal{X}_{n+h}(z)|\boldsymbol{\mathcal{X}}(z),\boldsymbol{\mathcal{B}}] = \widehat{\mu}(z) + \sum_{k=1}^{K} \widehat{\beta}_{n+h|n,k}\widehat{\phi}_k(z),$$

where $\widehat{\beta}_{n+h|n,k}$ represents the time series forecasts of the k^{th} principal component scores, which can be obtained by using a univariate time series forecasting method.

3.4 A univariate time series forecasting method

Hyndman and Shang, 2009 considered a univariate time series forecasting method to obtain $\widehat{\beta}_{n+h|n,k}$, such as autoregressive integrated moving average (ARIMA) model. This univariate time series forecasting method is able to model non-stationary time series containing a stochastic trend component. Since the yearly age-specific mortality rates do not contain seasonality, the ARIMA has a general form of

$$(1-\psi_1 B - \cdots - \psi_p B^p)(1-B)^d \boldsymbol{\beta}_k = \alpha + (1+\theta_1 B + \cdots + \theta_q B^q)\boldsymbol{w}_k,$$

where α represents the intercept, $(\psi_1,\dots,\psi_p)$ denote the coefficients associated with the autoregressive component, $(\theta_1,\dots,\theta_q)$ denote the coefficients associated with the moving average component, B denotes the backshift operator, d denotes the differencing operator, and

$\boldsymbol{w}_k = \{w_{1,k}, \dots, w_{n,k}\}$ represents a white-noise error term. We use the automatic algorithm of Hyndman and Khandakar, 2008 to choose the optimal orders of autoregressive p, moving average q and difference order d. The value of d is selected based on successive Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit-root tests (Kwiatkowski et al., 1992). KPSS tests are used for testing the null hypothesis that an observable time series is stationary around a deterministic trend. We first test the original time series for a unit root; if the test result is significant, then we test the differenced time series for a unit root. The procedure continues until we obtain our first insignificant result. Having determined d, the orders of p and q are selected based on the optimal Akaike information criterion (AIC) with a correction for small sample sizes (Akaike, 1974; Hurvich and Tsai, 1989). Having identified the optimal ARIMA model, maximum likelihood method can then be used to estimate the parameters.

4 Grouped functional time series forecasting techniques

4.1 Notation

For ease of explanation, we will introduce the notation using the Japanese example. The generalization to other contexts should be apparent. The Japanese data follow a multi-level geographical hierarchy coupled with a sex grouping variable. The geographical hierarchy is shown in Figure 3. Japan is split into eight regions, which in turn can be split into 47 prefectures.

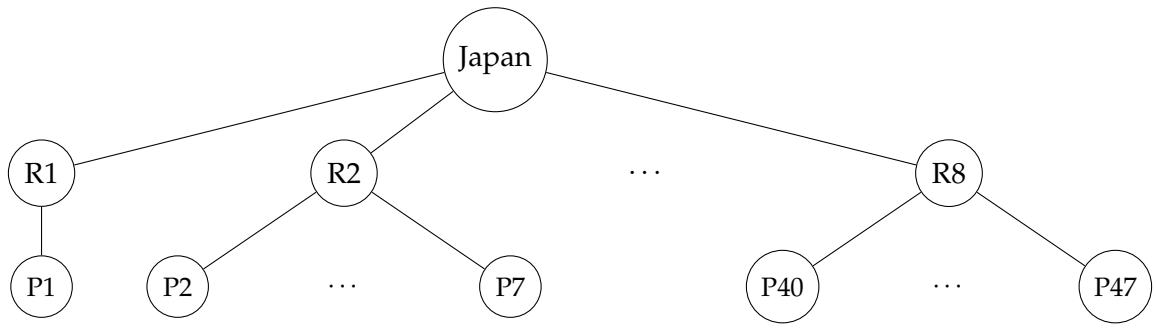


Figure 3: *The Japanese geographical hierarchy tree diagram, with eight regions and 47 prefectures.*

The data can also be split by sex. So each of the nodes in the geographical hierarchy can also be split into both males and females. We refer to a particular disaggregated series using the notation X*S meaning the geographical area X and the sex S, where X can take the values shown in Figure 3 and S can take values M (males), F (females) or T (total). For example: R1*F denotes

females in Region 1; P1*T denotes females and males in Prefecture 1; Japan*M denotes males in Japan; and so on.

Let $E_{\text{X*S},t}(z)$ denote the exposure-at-risk for series X*S in year t and age z, and let $D_{\text{X*S},t}(z)$ be the number of deaths for series X*S in year t and age z. Then the age-specific mortality rate is given by $R_{\text{X*S},t}(z) = D_{\text{X*S},t}(z)/E_{\text{X*S},t}(z)$. To simplify expressions, we will drop the age argument (z). Then for a given age, we can write

$$
\underbrace{\begin{bmatrix}
R_{\text{Japan*T},t}\\
R_{\text{Japan*F},t}\\
R_{\text{Japan*M},t}\\
R_{\text{R1*T},t}\\
R_{\text{R2*T},t}\\
\vdots\\
R_{\text{R8*T},t}\\
R_{\text{R1*F},t}\\
R_{\text{R2*F},t}\\
\vdots\\
R_{\text{R8*F},t}\\
R_{\text{R1*M},t}\\
R_{\text{R2*M},t}\\
\vdots\\
R_{\text{R8*M},t}\\
R_{\text{P1*T},t}\\
R_{\text{P2*T},t}\\
\vdots\\
R_{\text{P47*T},t}\\
R_{\text{P1*F},t}\\
R_{\text{P1*M},t}\\
R_{\text{P2*F},t}\\
R_{\text{P2*M},t}\\
\vdots\\
R_{\text{P47*F},t}\\
R_{\text{P47*M},t}
\end{bmatrix}}_{\boldsymbol{R}_t}
=
\underbrace{\begin{bmatrix}
\frac{E_{\text{P1*F},t}}{E_{\text{Japan*T},t}} & \frac{E_{\text{P1*M},t}}{E_{\text{Japan*T},t}} & \frac{E_{\text{P2*F},t}}{E_{\text{Japan*T},t}} & \frac{E_{\text{P2*M},t}}{E_{\text{Japan*T},t}} & \frac{E_{\text{P3*F},t}}{E_{\text{Japan*T},t}} & \frac{E_{\text{P3*M},t}}{E_{\text{Japan*T},t}} & \cdots & \frac{E_{\text{P47*F},t}}{E_{\text{Japan*T},t}} & \frac{E_{\text{P47*M},t}}{E_{\text{Japan*T},t}}\\
\frac{E_{\text{P1*F},t}}{E_{\text{Japan*F},t}} & 0 & \frac{E_{\text{P2*F},t}}{E_{\text{Japan*F},t}} & 0 & \frac{E_{\text{P3*F},t}}{E_{\text{Japan*F},t}} & 0 & \cdots & \frac{E_{\text{P47*F},t}}{E_{\text{Japan*F},t}} & 0\\
0 & \frac{E_{\text{P1*M},t}}{E_{\text{Japan*M},t}} & 0 & \frac{E_{\text{P2*M},t}}{E_{\text{Japan*M},t}} & 0 & \frac{E_{\text{P3*M},t}}{E_{\text{Japan*M},t}} & \cdots & 0 & \frac{E_{\text{P47*M},t}}{E_{\text{Japan*M},t}}\\
\frac{E_{\text{P1*F},t}}{E_{\text{R1,T},t}} & \frac{E_{\text{P1*M},t}}{E_{\text{R1,T},t}} & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & \frac{E_{\text{P2*F},t}}{E_{\text{R2,T},t}} & \frac{E_{\text{P2*M},t}}{E_{\text{R2,T},t}} & \frac{E_{\text{P3*F},t}}{E_{\text{R2,T},t}} & \frac{E_{\text{P3*M},t}}{E_{\text{R2,T},t}} & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & \frac{E_{\text{P47*F},t}}{E_{\text{R8,T},t}} & \frac{E_{\text{P47*M},t}}{E_{\text{R8,T},t}}\\
\frac{E_{\text{P1*F},t}}{E_{\text{R1,F},t}} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & \frac{E_{\text{P2*F},t}}{E_{\text{R2,F},t}} & 0 & \frac{E_{\text{P3*F},t}}{E_{\text{R2,F},t}} & 0 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & \frac{E_{\text{P47*F},t}}{E_{\text{R8,F},t}} & 0\\
0 & \frac{E_{\text{P1*M},t}}{E_{\text{R1,M},t}} & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & 0 & \frac{E_{\text{P2*M},t}}{E_{\text{R2,M},t}} & 0 & \frac{E_{\text{P3*M},t}}{E_{\text{R2,M},t}} & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \frac{E_{\text{P47*M},t}}{E_{\text{R8,M},t}}\\
\frac{E_{\text{P1*F},t}}{E_{\text{P1,T},t}} & \frac{E_{\text{P1*M},t}}{E_{\text{P1,T},t}} & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & \frac{E_{\text{P2*F},t}}{E_{\text{P2,T},t}} & \frac{E_{\text{P2*M},t}}{E_{\text{P2,T},t}} & 0 & 0 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & \frac{E_{\text{P47*F},t}}{E_{\text{P47,T},t}} & \frac{E_{\text{P47*M},t}}{E_{\text{P47,T},t}}\\
1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}}_{\boldsymbol{S}_t}
\underbrace{\begin{bmatrix}
R_{\text{P1*F},t}\\
R_{\text{P1*M},t}\\
R_{\text{P2*F},t}\\
R_{\text{P2*M},t}\\
\vdots\\
R_{\text{P47*F},t}\\
R_{\text{P47*M},t}
\end{bmatrix}}_{\boldsymbol{b}_t}
$$

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

Hyndman, Ahmed, et al., 2011 considered four hierarchical forecasting methods for univariate time series, namely the top-down, bottom-up, middle-out and optimal combination methods. Among the four, only bottom-up and optimal combination methods are suitable for forecasting a non-unique group structure. These two methods are reviewed in Sections 4.2 and 4.3, and their point and interval forecast accuracy comparisons with the independent forecasting method are presented in Sections 5.2 and 6.2.

4.2 Bottom-up method

One of the commonly used methods to forecasting grouped time series is the bottom-up method (e.g., Dangerfield and Morris, 1992; Zellner and Tobias, 2000). This method involves first generating base forecasts for each of the most disaggregated series and then aggregating these to produce all required forecasts. For example, let us consider the Japanese data. We first generate h-step-ahead base forecasts for the most disaggregated series, namely $\widehat{\boldsymbol{b}}_{n+h} = \left[\widehat{R}_{\text{P1*F},n+h}, \widehat{R}_{\text{P1*M},n+h}, \widehat{R}_{\text{P2*F},n+h}, \widehat{R}_{\text{P2*M},n+h}, \dots, \widehat{R}_{\text{P47*F},n+h}, \widehat{R}_{\text{P47*M},n+h}\right]^\top$.

Then the historical ratios that form the $\boldsymbol{S}_t$ summing matrix are forecast using an automated ARIMA algorithm (Hyndman and Khandakar, 2008). That is, let $p_t = E_{\text{X*S},t}/E_{\text{Y*W},t}$ be a non-zero element of $\boldsymbol{S}_t$. We forecast each time series $\{p_1, \dots, p_n\}$ for h-step-ahead to obtain $\hat{p}_{n+h}$. These are then used to form the matrix $\boldsymbol{S}_{n+h}$. Thus we obtain reconciled forecasts for all series:

$$\overline{\boldsymbol{R}}_{n+h} = \boldsymbol{S}_{n+h}\widehat{\boldsymbol{b}}_{n+h}.$$

The bottom-up method has the agreeable feature that it is simple and intuitive, and always results in series that are "aggregate consistent" (i.e., that the resulting forecasts satisfy the same aggregation constraints as the original data). The method performs well when the signal-to-noise ratio is relatively strong for the most disaggregated series. On the other hand, it may lead to inaccurate forecasts of the top-level series, in particular when there are missing or noisy data at the bottom level (see for example, Schwarzkopf, Tersine, and Morris, 1988; Shlifer and Wolff, 1979, in the univariate time series context).

4.3 Optimal combination method

Instead of considering only the bottom-level series, Hyndman, Ahmed, et al. (2011) proposed a method in which base forecasts for all aggregated and disaggregated series are computed independently, and then the resulting forecasts are reconciled so that they satisfy the aggregation constraints. As the base forecasts are independently generated, they will not usually be "aggregate consistent". The optimal combination method combines the base forecasts through linear regression by generating a set of revised forecasts that are as close as possible to the base forecasts but that also aggregate consistently within the group. The method is derived by writing the base forecasts as the response variable of the linear regression

$$\widehat{\boldsymbol{R}}_{n+h} = \boldsymbol{S}_{n+h}\boldsymbol{\beta}_{n+h} + \boldsymbol{\varepsilon}_{n+h},$$

where $\widehat{\boldsymbol{R}}_{n+h}$ is a matrix of h-step-ahead base forecasts for all series, stacked in the same order as for original data; $\boldsymbol{\beta}_{n+h} = \mathrm{E}[\boldsymbol{b}_{n+h} \mid \boldsymbol{R}_1, \ldots, \boldsymbol{R}_n]$ is the unknown mean of the forecast distributions of the most disaggregated series; and $\boldsymbol{\varepsilon}_{n+h}$ represents the reconciliation errors.

To estimate the regression coefficients, Hyndman, Ahmed, et al., 2011 and Hyndman, Lee, and Wang (2016) proposed a weighted least squares solution which we adapt to our problem as follows:

$$\widehat{\boldsymbol{\beta}}_{n+h} = \left(\boldsymbol{S}_{n+h}^{\top}\boldsymbol{W}^{-1}\boldsymbol{S}_{n+h}\right)^{-1}\boldsymbol{S}_{n+h}^{\top}\boldsymbol{W}^{-1}\widehat{\boldsymbol{R}}_{n+h},$$

where $\boldsymbol{W}$ is a diagonal matrix containing the one-step-ahead forecast variances for each series. Then the revised forecasts are given by

$$\overline{\boldsymbol{R}}_{n+h} = \boldsymbol{S}_{n+h}\widehat{\boldsymbol{\beta}}_{n+h} = \boldsymbol{S}_{n+h}\left(\boldsymbol{S}_{n+h}^{\top}\boldsymbol{S}_{n+h}\right)^{-1}\boldsymbol{S}_{n+h}^{\top}\widehat{\boldsymbol{R}}_{n+h}.$$

By construction, these are aggregate consistent and involve a combination of all the base forecasts. They are also unbiased since $\mathrm{E}[\overline{\boldsymbol{R}}_{n+h}] = \boldsymbol{S}_{n+h}\boldsymbol{\beta}_{n+h}$.

4.4 Constructing uniform and pointwise prediction intervals

To assess the forecast uncertainty, we adapt the method of Aue, Norinho, and Hörmann, 2015 for computing uniform and pointwise prediction intervals. The method can be summarized in the following steps:

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

$$\mathcal{X}_{\zeta+h}(z) = \widehat{\beta}_{\zeta+h,1}\widehat{\phi}_1(z) + \cdots + \widehat{\beta}_{\zeta+h,K}\widehat{\phi}_K(z),$$

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

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

$$\widehat{\epsilon}_{\omega}(z) = \mathcal{X}_{\zeta+h}(z) - \widehat{\mathcal{X}}_{\zeta+h}(z),$$

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

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 $\gamma^l(z)$ and $\gamma^u(z)$, respectively. We then seek a tuning parameter φ_α such that $\alpha \times 100\%$ of the residual functions satisfy

$$\varphi_\alpha \times \gamma^l(z) \leq \widehat{\epsilon}_{\omega}(z) \leq \varphi_\alpha \times \gamma^u(z), \qquad z \in \mathcal{I}.$$

   The residuals $\widehat{\epsilon}_1(z),\dots,\widehat{\epsilon}_M(z)$ are expected to be approximately stationary and, by the law of large numbers, to satisfy

$$\frac{1}{M}\sum_{\omega=1}^{M} \mathbb{1}\left(\varphi_\alpha \times \gamma^l(z) \leq \widehat{\epsilon}_{\omega}(z) \leq \varphi_\alpha \times \gamma^u(z)\right) \approx \Pr\left[\varphi_\alpha \times \gamma^l(z) \leq \mathcal{X}_{n+h}(z) - \widehat{\mathcal{X}}_{n+h}(z) \leq \varphi_\alpha \times \gamma^u(z)\right].$$

Note that Aue, Norinho, and Hörmann, 2015 calculate the standard deviation of $[\widehat{\epsilon}_1(z),\dots,\widehat{\epsilon}_M(z)]$, which leads to a parametric approach of constructing prediction intervals. Here we consider a nonparametric approach, as it allows us to reconcile bootstrapped forecasts among different functional time series in a hierarchy. Step 3 can easily be extended to pointwise prediction interval, where we determine a tuning parameter π_α such that $\alpha \times 100\%$ of the residual data points satisfy

$$\pi_\alpha \times \gamma^l(z_j) \leq \widehat{\epsilon}_\omega(z_j) \leq \pi_\alpha \times \gamma^u(z_j),$$

where j symbolizes discretized data points. Then, the h-step-ahead pointwise prediction intervals are given as

$$\pi_\alpha \times \gamma^l(z_j) \leq \mathcal{X}_{n+h}(z_j) - \widehat{\mathcal{X}}_{n+h}(z_j) \leq \pi_\alpha \times \gamma^u(z_j).$$

5 Results of the point forecasts

5.1 Point forecast evaluation

A rolling window analysis of a time series model is commonly used to assess model and parameter stabilities over time. It assesses the constancy of a model's parameter by computing parameter estimates and their forecasts over a rolling window of a fixed size through the sample (see Zivot and Wang, 2006, Chapter 9 for details). Using the first 29 observations from 1975 to 2003 in the Japanese age-specific mortality rates, we produce one- to ten-step-ahead point forecasts. Through a rolling window approach, we re-estimate the parameters in the univariate time series forecasting models using the first 30 observations from 1975 to 2004. Forecasts from the estimated models are then produced for one to nine-step-ahead. We iterate this process by increasing the sample size by one year until reaching the end of data period in 2013. This process produces 10 one-step-ahead forecasts, 9 two-step-ahead forecasts, ..., and 1 ten-step-ahead forecast. We compare these forecasts with the holdout samples to determine the out-of-sample point forecast accuracy.

To evaluate the point forecast accuracy, we use the mean absolute forecast error (MAFE) and root mean squared forecast error (RMSFE). They measure how close the forecasts are in comparison

to the actual values of the variable being forecast. For each series k, and they can be written as

$$\text{MAFE}_k(h) = \frac{1}{101\times(11-h)}\sum_{\varsigma=h}^{10}\sum_{j=1}^{101}\left|\mathcal{X}_{n+\varsigma}^{k}(z_j) - \widehat{\mathcal{X}}_{n+\varsigma}^{k}(z_j)\right|,$$

$$\text{RMSFE}_k(h) = \sqrt{\frac{1}{101\times(11-h)}\sum_{\varsigma=h}^{10}\sum_{j=1}^{101}\left[\mathcal{X}_{n+\varsigma}^{k}(z_j) - \widehat{\mathcal{X}}_{n+\varsigma}^{k}(z_j)\right]^2},$$

where $\mathcal{X}_{n+\varsigma}^{k}(z_j)$ represents the actual holdout sample for the j^{th} age and ς^{th} curve of the forecasting period in the k^{th} series, while $\widehat{\mathcal{X}}_{n+\varsigma}^{k}(z_j)$ represents the point forecasts for the holdout sample.

By averaging $\text{MAFE}_k(h)$ and $\text{RMSFE}_k(h)$ across the number of series within each level of disaggregation, we obtain an overall assessment of the point forecast accuracy for each level within the collection of series, denoted by $\text{MAFE}(h)$ and $\text{RMSFE}(h)$. They are defined as

$$\text{MAFE}(h) = \frac{1}{m_k}\sum_{k=1}^{m_k}\text{MAFE}_k(h),$$

$$\text{RMSFE}(h) = \frac{1}{m_k}\sum_{k=1}^{m_k}\text{RMSFE}_k(h),$$

where m_k denotes the number of series at the k^{th} level of disaggregation, for $k = 1,\dots,K$.

For 10 different forecast horizons, we consider two summary statistics to evaluate point forecast accuracy between the methods for national and sub-national population. The summary statistics chosen are the mean and median values due to their suitability for handling squared and absolute errors (Gneiting, 2011). They are given by

$$\text{Mean}\,(\text{RMSFE}) = \frac{1}{10}\sum_{h=1}^{10}\text{RMSFE}(h),$$

$$\text{Median}\,(\text{MAFE}) = \frac{1}{2}\left[\text{MAFE}(5) + \text{MAFE}(6)\right],$$

where $[5]$ and $[6]$ represent the 5^{th} and 6^{th} terms after ranking $\text{MAFE}(h)$ for $h = 1,2,\dots,10$ from smallest to largest.

5.2 Point forecast comparison

Averaging over all series at each level of the Japanese data hierarchy, Tables 2 and 3 present MAFE(h) and RMSFE(h) values using the independent functional time series and two grouped functional time series forecasting methods. The bold entries highlight the method that performs the best for each level of the hierarchy and each forecast horizon, based on the smallest forecast error. In the short-term forecast horizon, the independent functional time series forecasting and optimal combination methods generally have the smaller forecast errors than the bottom-up method. As the forecast horizon increases from $h = 3$ to $h = 10$, the bottom-up method performs the best with the smallest forecast errors. At the bottom level, it is not surprising that the independent functional time series and bottom-up methods produce the same forecast accuracy. Averaged over all levels of a hierarchy, it is advantageous to use the grouped functional time series forecasting methods over the independent functional time series forecasting method. For this example, we recommend the bottom-up method.

Table 2: *MAFEs ($\times$100) in the holdout sample between the independent functional time series forecasting and two grouped functional time series forecasting methods applied to the Japanese age-specific mortality rates. The bold entries highlight the method that performs best for each level of the hierarchy and each forecast horizon, as well as summary statistic.*

Forecasting method	h	Total	Sex	Region	Region (Sex)	Prefecture	Prefecture (Sex)
Independent	1	0.134	0.133	**0.157**	0.209	0.252	0.378
	2	0.194	0.181	0.189	**0.225**	**0.253**	0.390
	3	0.220	0.213	0.212	0.235	0.263	**0.365**
	4	0.256	0.259	0.248	0.262	0.279	**0.374**
	5	0.290	0.301	0.272	0.287	0.292	**0.381**
	6	0.323	0.334	0.300	0.312	0.311	**0.399**
	7	0.375	0.393	0.347	0.357	0.337	**0.420**
	8	0.415	0.432	0.388	0.398	0.367	**0.445**
	9	0.461	0.460	0.412	0.411	0.378	**0.451**
	10	0.457	0.427	0.395	0.391	0.366	**0.437**
	Median	0.306	0.318	0.286	0.299	0.301	**0.394**
Bottom-up	1	0.116	0.134	0.179	0.220	0.256	0.378
	2	0.123	**0.142**	0.196	0.235	0.273	0.390
	3	**0.129**	**0.151**	**0.166**	**0.216**	**0.242**	**0.365**
	4	**0.142**	**0.178**	**0.177**	**0.234**	**0.248**	**0.374**
	5	**0.138**	**0.202**	**0.178**	**0.249**	**0.249**	**0.381**
	6	**0.160**	**0.234**	**0.192**	**0.273**	**0.260**	**0.399**
	7	**0.179**	**0.283**	**0.211**	**0.313**	**0.268**	**0.420**
	8	**0.205**	**0.322**	**0.236**	**0.354**	**0.283**	**0.445**
	9	**0.228**	**0.353**	**0.248**	**0.371**	**0.283**	**0.451**

Continued on next page

Forecasting method	h	Total	Sex	Region	Region (Sex)	Prefecture	Prefecture (Sex)
	10	**0.209**	**0.329**	**0.231**	**0.354**	**0.267**	**0.437**
	Median	**0.151**	**0.218**	**0.194**	**0.261**	**0.264**	**0.394**
Optimal combination	1	**0.111**	**0.130**	0.164	**0.207**	**0.247**	**0.371**
	2	**0.120**	0.149	**0.181**	0.226	0.261	**0.383**
	3	0.139	0.176	0.168	0.224	0.246	0.373
	4	0.164	0.217	0.190	0.255	0.258	0.388
	5	0.183	0.258	0.203	0.284	0.266	0.404
	6	0.208	0.293	0.223	0.314	0.280	0.426
	7	0.248	0.352	0.255	0.364	0.299	0.456
	8	0.280	0.394	0.291	0.413	0.321	0.487
	9	0.301	0.422	0.302	0.427	0.326	0.497
	10	0.282	0.399	0.286	0.412	0.310	0.483
	Median	0.195	0.276	0.213	0.299	0.273	0.415

Table 3: *RMSFEs ($\times$100) in the holdout sample between the independent functional time series forecasting and two grouped functional time series forecasting methods applied to the Japanese age-specific mortality rates. The bold entries highlight the method that performs best for each level of the hierarchy and each forecast horizon, as well as summary statistic.*

Forecasting method	h	Total	Sex	Region	Region (Sex)	Prefecture	Prefecture (Sex)
Independent	1	0.468	**0.464**	**0.528**	0.719	**0.812**	1.300
	2	0.589	0.573	**0.611**	**0.765**	**0.814**	1.367
	3	0.658	0.657	0.680	0.804	0.843	**1.235**
	4	0.740	0.776	0.765	0.880	0.885	**1.264**
	5	0.812	0.876	0.824	0.951	0.917	**1.285**
	6	0.876	0.946	0.876	0.996	0.958	**1.320**
	7	0.992	1.087	0.982	1.117	1.021	**1.375**
	8	1.084	1.176	1.068	1.205	1.077	**1.418**
	9	1.170	1.222	1.101	1.210	1.084	**1.399**
	10	1.135	1.107	1.042	1.127	1.024	**1.331**
	Mean	0.852	0.888	0.848	0.977	0.943	**1.330**
Bottom up	1	**0.413**	0.469	0.614	0.740	0.856	1.300
	2	**0.423**	**0.495**	0.729	0.836	0.956	1.367
	3	**0.466**	**0.549**	**0.570**	**0.742**	**0.778**	**1.235**
	4	**0.513**	**0.624**	**0.613**	**0.800**	**0.804**	**1.264**
	5	**0.540**	**0.692**	**0.637**	**0.854**	**0.812**	**1.285**
	6	**0.579**	**0.750**	**0.671**	**0.900**	**0.840**	**1.320**
	7	**0.643**	**0.865**	**0.736**	**1.011**	**0.875**	**1.375**
	8	**0.706**	**0.948**	**0.794**	**1.099**	**0.910**	**1.418**
	9	**0.744**	**1.000**	**0.815**	**1.116**	**0.907**	**1.399**
	10	**0.673**	**0.899**	**0.752**	**1.038**	**0.842**	**1.331**
	Mean	**0.570**	**0.729**	**0.693**	**0.914**	**0.858**	**1.330**
Optimal combination	1	0.430	0.490	0.571	**0.712**	0.816	**1.276**
	2	0.462	0.546	0.654	0.795	0.881	**1.327**
	3	0.527	0.619	0.606	0.782	0.805	1.265
	4	0.592	0.714	0.666	0.863	0.843	1.307

Continued on next page

Forecasting method	h	Total	Sex	Region	Region (Sex)	Prefecture	Prefecture (Sex)
	5	0.644	0.805	0.710	0.939	0.867	1.343
	6	0.694	0.875	0.754	0.996	0.901	1.387
	7	0.779	1.004	0.839	1.122	0.957	1.459
	8	0.851	1.094	0.913	1.220	1.004	1.513
	9	0.889	1.146	0.926	1.234	1.003	1.497
	10	0.819	1.048	0.865	1.155	0.936	1.427
	Mean	0.669	0.834	0.750	0.982	0.901	1.380

6 Results of the interval forecasts

6.1 Interval forecast evaluation

In order to evaluate pointwise interval forecast accuracy, we utilize 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 were calculated at the $100(1-\alpha)\%$ nominal coverage probability. We consider the common case of the symmetric $100(1-\alpha)\%$ prediction interval, with lower and upper bounds that are predictive quantiles at $\alpha/2$ and $1-\alpha/2$, denoted by $\widehat{\mathcal{X}}^{l}_{n+h}(z_j)$ and $\widehat{\mathcal{X}}^{u}_{n+h}(z_j)$. As defined by Gneiting and Raftery, 2007, a scoring rule for the pointwise interval forecast at time point z_j is

$$\begin{aligned} S_\alpha\left[\widehat{\mathcal{X}}^{l}_{n+h}(z_j),\widehat{\mathcal{X}}^{u}_{n+h}(z_j);\mathcal{X}_{n+h}(z_j)\right] = &\left[\widehat{\mathcal{X}}^{u}_{n+h}(z_j)-\widehat{\mathcal{X}}^{l}_{n+h}(z_j)\right]+ \\ &\frac{2}{\alpha}\left[\widehat{\mathcal{X}}^{l}_{n+h}(z_j)-\mathcal{X}_{n+h}(z_j)\right]\mathbb{1}\left\{\mathcal{X}_{n+h}(z_j)<\widehat{\mathcal{X}}^{l}_{n+h}(z_j)\right\}+ \\ &\frac{2}{\alpha}\left[\mathcal{X}_{n+h}(z_j)-\widehat{\mathcal{X}}^{u}_{n+h}(z_j)\right]\mathbb{1}\left\{\mathcal{X}_{n+h}(z_j)>\widehat{\mathcal{X}}^{u}_{n+h}(z_j)\right\}, \end{aligned}$$

where α denotes the level of significance, customarily $\alpha=0.2$. The interval score rewards a narrow prediction interval, if and only if the true observation lies within the prediction interval. The optimal interval score is achieved when $\mathcal{X}_{n+h}(z_j)$ lies between $\widehat{\mathcal{X}}^{l}_{n+h}(z_j)$ and $\widehat{\mathcal{X}}^{u}_{n+h}(z_j)$, and the distance between $\widehat{\mathcal{X}}^{l}_{n+h}(z_j)$ and $\widehat{\mathcal{X}}^{u}_{n+h}(z_j)$ is minimal.

For different time points in a curve and different days in the forecasting period, the mean interval score is defined by

$$\overline{S}_\alpha(h)=\frac{1}{101\times(11-h)}\sum_{\varsigma=h}^{10}\sum_{j=1}^{101}S_\alpha\left[\widehat{\mathcal{X}}^{l}_{n+\varsigma}(z_j),\widehat{\mathcal{X}}^{u}_{n+\varsigma}(z_j);\mathcal{X}_{n+\varsigma}(z_j)\right],$$

where $S_\alpha\left[\widehat{\mathcal{X}}^l_{n+\varsigma}(z_j), \widehat{\mathcal{X}}^u_{n+\varsigma}(z_j); \mathcal{X}_{n+\varsigma}(z_j)\right]$ denotes the interval score at the ς^{th} curve of the forecasting period.

For 10 different forecast horizons, we consider two summary statistics to evaluate interval forecast accuracy. The summary statistics chosen are the mean and median values, given by

$$\text{Mean}\left(\overline{S}_\alpha\right) = \frac{1}{10}\sum_{h=1}^{10}\overline{S}_\alpha(h),$$
$$\text{Median}\left(\overline{S}_\alpha\right) = \frac{1}{2}\left[\overline{S}_\alpha(5) + \overline{S}_\alpha(6)\right].$$

6.2 Interval forecast comparison

In Table 4, we present the mean interval scores for one-step-ahead to ten-step-ahead forecasts, using the independent and two grouped functional time series forecasting methods. The independent functional time series generally gives the most accurate interval forecasts at the national level, while the grouped functional time series forecasting methods demonstrate superior forecast accuracy for the sub-national level. The bottom-up method gives the most accurate interval forecasts at the region level, while the optimal combination method gives the most accurate interval forecasts at the prefecture level. Based on the overall mean interval scores, the bottom-up methods outperform the independent functional time series forecasting and optimal combination methods, in terms of interval forecast accuracy. Thus, the bottom-up method is recommended for this example.

Table 4: *Mean interval scores ($\times$100) in the holdout sample between the independent functional time series forecasting and two grouped functional time series forecasting methods applied to the Japanese age-specific mortality rates. The bold entries highlight the method that performs best for each level of the hierarchy and each forecast horizon, as well as two summary statistics.*

Forecasting method	h	Total	Sex	Region	Region (Sex)	Prefecture	Prefecture (Sex)
Independent	1	**0.523**	**0.627**	1.396	2.184	**1.396**	2.184
	2	**0.676**	**0.819**	1.372	2.276	**1.372**	2.276
	3	**0.738**	**0.914**	1.373	2.060	**1.373**	2.060
	4	**0.910**	**1.076**	1.421	2.091	1.421	2.091
	5	**1.123**	**1.249**	1.484	2.104	1.484	2.104
	6	**1.198**	**1.315**	1.565	2.175	1.565	2.175
	7	**1.322**	**1.557**	1.643	2.250	1.643	2.250
	8	**1.390**	**1.666**	1.734	2.350	1.734	2.350
	9	**1.558**	**1.720**	1.754	2.294	1.754	2.294

Continued on next page

Forecasting method	h	Total	Sex	Region	Region (Sex)	Prefecture	Prefecture (Sex)
	10	1.437	**1.580**	1.752	2.307	1.752	2.307
	Mean	**1.088**	**1.252**	1.549	2.209	1.549	2.209
	Median	**1.160**	**1.282**	1.524	2.217	1.524	2.217
Bottom up	1	0.856	0.832	**0.974**	1.166	1.439	2.184
	2	0.972	0.955	**1.156**	1.321	1.578	2.276
	3	1.060	1.079	**0.885**	**1.131**	1.291	2.060
	4	1.206	1.268	**0.962**	**1.207**	1.316	2.091
	5	1.248	1.406	**0.978**	**1.261**	1.334	2.104
	6	1.354	1.584	**1.056**	**1.370**	1.372	2.175
	7	1.454	1.859	**1.120**	**1.508**	**1.414**	2.250
	8	1.538	2.067	**1.194**	**1.664**	**1.471**	2.350
	9	1.616	2.166	**1.185**	**1.698**	**1.409**	2.294
	10	**1.384**	1.881	**1.056**	**1.562**	1.449	2.307
	Mean	1.269	1.510	**1.057**	**1.389**	1.407	2.209
	Median	1.301	1.495	**1.056**	**1.346**	1.412	2.217
Optimal combination	1	0.924	0.861	0.995	**1.101**	**1.268**	**2.029**
	2	1.037	1.032	1.157	**1.247**	**1.363**	**2.089**
	3	1.241	1.270	1.066	1.208	**1.205**	**1.968**
	4	1.450	1.560	1.208	1.354	**1.251**	**2.014**
	5	1.582	1.809	1.289	1.494	**1.289**	**2.054**
	6	1.759	2.063	1.401	1.656	**1.344**	**2.122**
	7	1.957	2.441	1.556	1.889	1.416	**2.208**
	8	2.108	2.712	1.705	2.107	1.493	**2.315**
	9	2.207	2.846	1.719	2.166	1.468	**2.284**
	10	1.879	2.490	1.533	1.981	**1.429**	**2.299**
	Mean	1.614	1.908	1.363	1.620	**1.353**	**2.138**
	Median	1.670	1.936	1.345	1.575	**1.354**	**2.105**

7 Conclusion

We have extended two grouped time series forecasting methods, namely the bottom-up and optimal combination methods, from univariate to functional time series. These grouped functional time series forecasting methods were derived by coupling grouped univariate time series forecasting methods with functional time series analysis.

The bottom-up method models and forecasts data series at the most disaggregated level, and then aggregates the results using the summing matrix. In that summing matrix, each element is forecast from the historical data using univariate time series models.

The optimal combination method combines the base forecasts obtained from independent functional time series forecasting methods using linear regression. It generates a set of revised

forecasts that are as close as possible to the base forecasts, but that also aggregate consistently with the known grouping structure. Under some mild assumptions, the regression coefficient can be estimated by ordinary least squares.

Using age-specific mortality rates at the national and sub-national levels in Japan, we compare the one-step-ahead to ten-step-ahead forecast accuracy between the independent functional time series forecasting method and the two proposed grouped functional time series forecasting methods. We found that the grouped functional time series forecasting methods produced more accurate point and interval forecasts than those obtained by the independent functional time series forecasting method. In addition, the grouped functional time series forecasting methods produce forecasts that obey the natural group structure, thus giving forecast mortality rates at the sub-national levels that add up to the forecast mortality rates at the national level.

We have also presented a way of constructing uniform and pointwise prediction intervals for grouped functional time series using bootstrapping. The method calculates in-sample forecast errors between the in-sample holdout data and their reconstruction by functional principal component regression. By sampling with replacement from the bootstrapped in-sample errors, we obtain lower and upper bounds, and then find an optimal tuning parameter for achieving uniform or pointwise nominal coverage probability. With this tuning parameter, out-of-sample uniform or pointwise prediction intervals are obtained.

There are a few ways in which the paper can be further extended and we briefly outline three. First, the methodology can be applied to cause-specific mortality, considered in Murray and Lopez, 1997, Girosi and King, 2008 and Gaille and Sherris, 2015. Second, due to the availability of data, we have considered disaggregation of mortality by sex and geography. However, mortality rates can be further disaggregated with the inclusion of other factors, such as socioeconomic status *inter alia* (Bassuk, Berkman, and Amick III, 2002; Singh et al., 2013). Finally, coherent forecasting methods can be used to jointly model and forecast age-specific mortality rates from two or more populations (see for example, Hyndman, Booth, and Yasmeen, 2013; Li and Lee, 2005).

SUPPLEMENTARY MATERIAL

R-package for functional time series forecasting The R-package *ftsa* containing code to produce point and interval forecasts from independent functional time series forecasting method described in the article. The R-package can be obtained from CRAN (https://cran.r-project.org/web/packages/ftsa/index.html).

Code for grouped functional time series forecasting The R code to produce point and interval forecasts from the two grouped functional time series forecasts described in the article. (gfts.zip)

Code for shiny application The R code to produce a shiny user interface for plotting every series in the Japanese data hierarchy. (shiny.zip)

References

Akaike, H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Aue, A, DD Norinho, and S Hörmann (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* **110**(509), 378–392.

Bassuk, SS, LF Berkman, and BC Amick III (2002). Socioeconomic status and mortality among the elderly: Findings from four US communities. *American Journal of Epidemiology* **155**(6), 520–533.

Besse, P, H Cardot, and D Stephenson (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27**(4), 673–687.

Bosq, D (2000). *Linear Processes in Function Spaces*. New York: Lecture notes in Statistics.

Chang, W, J Cheng, J Allaire, Y Xie, and J McPherson (2015). *shiny: Web Application Framework for R*. R package version 0.12.2. %5Curl%7Bhttp://CRAN.R-project.org/package=shiny%7D.

Chiou, JM (2012). Dynamical functional prediction and classification with application to traffic flow prediction. *The Annals of Applied Statistics* **6**(4), 1588–1614.

Chiou, JM and HG Müller (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **104**(486), 572–585.

Chiou, JM, YF Yang, and YT Chen (2015). Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis* **in press**.

Coulmas, F (2007). *Population Decline and Ageing in Japan – the Social Consequences*. New York. Routledge.

D'Amato, V, G Piscopo, and M Russolillo (2011). The mortality of the Italian population: Smoothing technique on the Lee-Carter model. *The Annals of Applied Statistics* **5**(2A), 705–724.

Dangerfield, BJ and JS Morris (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting* **8**(2), 233–241.

de Boor, C (2001). *A Practical Guide to Splines*. Vol. 27. Applied Mathematical Sciences. New York: Springer.

Fang, L and WK Härdle (2015). *Stochastic population analysis: A functional data approach*. Working paper. Humboldt University of Berlin, Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2630301.

Gaille, SA and M Sherris (2015). Causes-of-death mortality: What do we know on their dependence? *North American Actuarial Journal* **19**(2), 116–128.

Girosi, F and G King (2008). *Demographic Forecasting*. Princeton: Princeton University Press.

Gneiting, T (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**(494), 746–762.

Gneiting, T and M Katzfuss (2014). Probabilistic forecasting. *The Annual Review of Statistics and Its Application* **1**, 125–151.

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

Hall, P (2011). "Principal component analysis for functional data: Methodology, theory, and discussion". In: *The Oxford Handbook of Functional Data Analysis*. Ed. by F Ferraty and Y Romain. New York: Oxford University Press, pp.210–234.

Hall, P and C Vial (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society (Series B)* **68**(4), 689–705.

Hörmann, S and P Kokoszka (2010). Weakly dependent functional data. *The Annals of Statistics* **38**(3), 1845–1884.

Hörmann, S and P Kokoszka (2012). "Functional Time Series". In: *Handbook of Statistics*. Ed. by TS Rao, SS Rao, and CR Rao. Vol. 30. Amsterdam: North Holland, pp.157–186.

Hurvich, CM and C Tsai (1989). Regression and time series model selection in small samples. *Biometrika* **76**(2), 297–307.

Hyndman, RJ, RA Ahmed, G Athanasopoulos, and HL Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis* **55**(9), 2579–2589.

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

Hyndman, RJ and HL Shang (2009). Forecasting functional time series (with discussions). *Journal of the Korean Statistical Society* **38**(3), 199–211.

Hyndman, RJ and HL Shang (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics* **19**(1), 29–45.

Hyndman, R and M Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis* **51**(10), 4942–4956.

Hyndman, RJ and Y Khandakar (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* **27**(3).

Hyndman, RJ, A Lee, and E Wang (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics and Data Analysis* **in press**.

Japanese Mortality Database (2015). *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/18/2015).

Kahn, KB (1998). Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting* **17**(2), 14–19.

Kokoszka, P and X Zhang (2012). Functional prediction of intraday cumulative returns. *Statistical Modelling* **12**(4), 377–398.

Kowal, DR, DS Matteson, and D Ruppert (2015). *A Bayesian multivariate functional dynamic linear model*. Tech. rep. Cornell University, Retrieved from http://arxiv.org/pdf/1411.0764.pdf.

Kwiatkowski, D, PCB Phillips, P Schmidt, and Y Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* **54**(1-3), 159–178.

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

Müller, HG (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32**(2), 223–240.

Murray, CJL and AD Lopez (1997). Alternative projections of mortality and disability by cause 1990-2020: Global burden of disease study. *The Lancet* **349**(9064), 1498–1504.

OECD (2013). *Pensions at a Glance 2013: OECD and G20 Indicators*. Tech. rep. OECD Publishing, Retrieved from http://dx.doi.org/10.1787/pension_glance-2013-en.

Peña, D and I Sánchez (2007). Measuring the advantages of multivariate vs univariate forecasts. *Journal of Time Series Analysis* **28**(6), 886–909.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. %5Curl%7Bhttp://www.R-project.org/%7D.

Ramsay, J and B Silverman (2005). *Functional Data Analysis*. 2nd. New York: Springer Series in Statistics.

Reiss, PT, J Goldsmith, HL Shang, and RT Ogden (2016). Methods for scalar-on-function regression. *International Statistical Review* **in press**.

Rice, J and B Silverman (1991). Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves. *Journal of the Royal Statistical Society (Series B)* **53**(1), 233–243.

Schwarzkopf, AB, RJ Tersine, and JS Morris (1988). Top-down versus bottom-up forecasting strategies. *International Journal of Production Research* **26**(11), 1833–1843.

Sefton, J and M Weale (2009). *Reconciliation of National Income and Expenditure: Balanced Estimates of National Income for the United Kingdom, 1920-1990*. Cambridge: Cambridge University Press.

Shang, HL (2014). A survey of functional principal component analysis. *AStA Advance in Statistical Analysis* **98**(2), 121–142.

Shang, HL and RJ Hyndman (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation* **81**(7), 1310–1324.

Shlifer, E and RW Wolff (1979). Aggregation and proration in forecasting. *Management Science* **25**(6), 594–603.

Singh, GK, RE Azuine, M Siahpush, and MD Kogan (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.

Suzuki, M, B Willcox, and C Willcox (2004). Successful aging: Secrets of Okinawan longevity. *Geriatrics & Gerontology International* **4**(s1), S180–S181.

Takata, H, T Ishii, M Suzuki, S Sekiguchi, and H Iri (1987). Influence of major histocompatibility complex region genes on human longevity among Okinawan-Japanese centenarians and nonagenarians. *The Lancet* **330**(8563), 824–826.

Wang, JL, JM Chiou, and HG Müller (2015). *Review of functional data analysis*. Working paper. University of California, Davis, Retrieved from http://arxiv.org/pdf/1507.05135v1.pdf.

Willcox, DC, BJ Willcox, S Shimajiri, S Kurechi, and M Suzuki (2007). Aging gracefully: A retrospective analysis of functional status in Okinawan centenarians. *The American Journal of Geriatric Psychiatry* **15**(3), 252–256.

Yao, F, HG Müller, and JL Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**(470), 577–590.

Zellner, A and J Tobias (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting* **19**(5), 457–469.

Zivot, E and J Wang (2006). *Modeling Financial Time Series with S-PLUS*. New York: Springer. Chap. Rolling analysis of time series.