International Journal of Forecasting 32 (2016) 629–649

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A multilevel functional data method for forecasting population, with an application to the United Kingdom

Han Lin Shang [a,d,*], Peter W.F. Smith [b], Jakub Bijak [a], Arkadiusz Wiśniowski [c]

[a] ESRC Centre for Population Change, University of Southampton, United Kingdom
[b] Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom
[c] Social Statistics Discipline Area, School of Social Sciences, University of Manchester, United Kingdom
[d] Research School of Finance, Actuarial Studies and Statistics, Australian National University, Australia

A R T I C L E I N F O

Keywords:
Age-specific population forecasting
Population projection matrix
Leslie matrix
Linear growth models
Functional data analysis
Functional principal component analysis

A B S T R A C T

Cohort component models are often used to model the evolution of an age-specific population, and are particularly useful for highlighting the demographic component that contributes the most to the population change. Recently, most of the attention has been devoted to the estimation of four specific demographic components, namely mortality, fertility, emigration and immigration. Many methods take a deterministic viewpoint, which can be quite restrictive in practice. The statistical method that we propose is a multilevel functional data method, where both mortality and migration are modelled and forecast jointly for females and males. The forecast uncertainty associated with each component is incorporated through parametric bootstrapping. Using the historical data for the United Kingdom from 1975 to 2009, we found that the proposed method shows a good in-sample forecast accuracy for the holdout data between 2001 and 2009. Moreover, we produce out-of-sample population forecasts from 2010 to 2030, and compare our forecasts with those produced by the Office for National Statistics.

1. Introduction

Recent decades have seen considerable developments in the stochastic modelling and forecasting of populations. This includes the pioneering work of Ahlburg and Land (1992), Alho and Spencer (1985, 2005), Alho et al. (2006), Bongaarts and Bulatao (2000), Bryant and Graham (2013), De Beer (2000), Hyndman and Booth (2008), Keilman (1990), Lutz (1996), Lee and Tuljapurkar (1994), Raftery, Li, Ševčiková, Gerland, and Heilig (2012), and Wiśniowski, Smith, Bijak, Raymer, and Forster (2015), among many others. Most of this body of work emphasises the advantages of stochastic modelling and forecasting over its deterministic counterparts (for reviews, see Booth, 2006; Raftery, Lalic, & Gerland, 2014; Wilson & Bell, 2007; Wilson & Rees, 2005). While the stochastic approach incorporates uncertainty into population estimates and forecasts, the deterministic approach provides scenarios that are often implausible, assuming high, medium and low rates, but lacking specificity regarding the probability ranges of these three variants. Despite the advantages of stochastic forecasts, they have not been used widely by official statistical agencies, for several reasons (Keilman, Pham, & Hetland, 2002; Lutz & Goldstein, 2004). First, it is not straightforward to consider all of the different types of forecast uncertainties. Second, some official statistical agencies are often actually constrained to the use of a

* Correspondence to: Research School of Finance, Actuarial Studies and Statistics, Australian National University, Kingsley Street, Canberra, ACT 0200, Australia. Tel.: +61 2 612 50535; fax: +61 2 612 50087.
E-mail address: hanlin.shang@anu.edu.au (H.L. Shang).

http://dx.doi.org/10.1016/j.ijforecast.2015.10.002
0169-2070/

limited set of statistical methods with which they are familiar for official forecasts (see also Wiśniowski et al., 2015). While much has been undertaken, a lot of work is still needed in order to produce usable probabilistic models.

This paper considers a functional data method for forecasting age- and sex-specific populations (see also Hyndman & Booth, 2008). As a generalisation of Lee and Carter's (1992) method, the functional data analysis views each component of the population from a functional perspective, by combining ideas from nonparametric smoothing (Eubank, 1999), functional principal component regression (Hyndman & Ullah, 2007), and bootstrapping (Efron, 2011). Each component of the population is smoothed using a tailored nonparametric smoothing technique, and is modelled as a continuous function of age, so that the patterns of variations among years are captured by the functional principal components and their associated scores. By drawing from normal distributions, the simulated scores can be forecast using a univariate time series technique for each replication. The probabilistic forecasts can be obtained through parametric bootstrapping by conditioning on the estimated mean and functional principal component functions.

Demographic processes, such as fertility, mortality, and migration, tend to exhibit strong regularities in their age patterns. Modelling age profiles over time permits a relatively concise representation of the history of demographic patterns. Time series methods can be utilised to extrapolate age profiles into future years. Throughout this paper, we focus on the exploration of independent functional data and multilevel functional data methods for forecasting age-specific fertility rates, mortality rates, emigration rates and immigration counts in a cohort component projection model. We also focus on the selection of optimal Box–Cox transformation parameters based on an in-sample measure of uncertainty for forecasting age-specific fertility, mortality, emigration and immigration in a cohort component projection model. As an illustration, we use a time series from the United Kingdom (UK), consisting of age- and sex-specific data for single year of age from 1975 to 2009. We use the term "forecast" to refer to an outcome of a probabilistic exercise in predictions, as opposed to a purely deterministic "projection" (Keilman, 1990).

The independent functional data method was first implemented by Hyndman and Booth (2008) for forecasting age- and sex-specific populations in Australia. Our work differs from that of Hyndman and Booth (2008) in four main ways. (1) We use the multilevel functional data method to model both mortality and migration for females and males jointly. This allows improved point and interval forecast accuracies to be obtained, as is evident from the in-sample population forecasts in Section 5. (2) We use a stochastic search algorithm to find the optimal Box–Cox transformation parameters for mortality, fertility and migration, based on a measure of forecast uncertainty (see also Shang, 2015). (3) We model the emigration and immigration separately, instead of considering the net migration as per Hyndman and Booth (2008), Raftery et al. (2012) and Wilson and Bell (2007). (4) Bootstrapping is our chosen method for constructing prediction intervals, and does not depend on the normality assumption.

This article is organised as follows. In Section 2, we present a general background to population forecasting, define the age- and sex-specific population projection matrix, and highlight the issues that pertain to the stochastic modelling and forecasting of each demographic component by age and sex. In Section 3, we introduce two functional data methods for forecasting the age profiles of each demographic component. In Section 4, we evaluate and compare the in-sample point and interval forecast accuracies using the UK data for 1975–2009, with a holdout sample from 2000 to 2009. In Section 5, we present out-of-sample forecasts from 2010 to 2030. Conclusions are provided in Section 6, along with some ideas as to possible extensions of the methods presented here.

2. Cohort component population projection model

We consider the cohort component model for describing the evolution of an age- and sex-specific population (see also Leslie, 1945; Preston, Heuveline, & Guillot, 2001; Rogers, 1995, pp. 117–137; Wiśniowski et al., 2015). For each sex and age or age group, we need to estimate mid-year fertility rates, mortality rates, emigration rates and immigration counts. We work with a single year of age. Instead of modelling the net migration, we choose to model emigration rates and immigration counts separately, because immigration does not have an easily defined population at risk; thus, we model the counts instead, which is a common practice in population forecasting (McDonald & Kippen, 2002; Raymer, Abel, & Rogers, 2012; Rees, 1986; Wiśniowski et al., 2015). First, the accurate estimation and forecasting of the mortality is important for the calculation of the survival rate s_i for ages $i = 0, \dots, 89, 90+$. The survival rate measures the proportion of the people in age group i who will survive to the beginning of the next period. Second, the fertility rate $f_{i,t}$ measures the yearly average number of live births to women aged i in year t. Because of the relatively small counts for very young and very old ages, the data on birth counts were aggregated into the age groups "under 15 years" and "45+ years". Third, the emigration rate η_i, for $i = 0, \dots, 89, 90+$, measures the average number of migrants at age i. Finally, the immigration count I_i measures the total number of immigrants at age i.

As per Preston et al. (2001) and Wiśniowski et al. (2015), let $\mathbf{b}_t^k = (0, \dots, b_{14,t}^k, b_{15,t}^k \dots, b_{44,t}^k, b_{45,t}^k, \dots, 0)$ be a vector of birth rates, which can be derived from the age-specific fertility rates as follows:

$$b_{i,t}^k = \frac{1}{1 + 0.5\mu_{0,t}^k} \times a_k \times \frac{1}{2}\left(f_{i,t} + s_{i,t}^{\mathrm{F}} f_{i+1,t}\right), \qquad (1)$$

where $f_{i,t} > 0$ represents the fertility rate at age i in year t; $\mu_{0,t}^k, k = \mathrm{F}$ or M, represents the age-specific female or male mortality rate at age 0; $a_{\mathrm{F}} = 1/(1 + 1.05)$ is the proportion of female births amongst all live births, and the sex ratio 1.05 is determined based on the numbers of live births in England and Wales in 2012 (Office for National Statistics,