J Pop Research (2015) 32:69–79
DOI 10.1007/s12546-014-9138-0

Selection of the optimal Box–Cox transformation parameter for modelling and forecasting age-specific fertility

Han Lin Shang

Published online: 4 December 2014

Abstract The Box–Cox transformation can sometimes yield noticeable improvements in model simplicity, variance homogeneity and precision of estimation, such as in modelling and forecasting age-specific fertility. Despite its importance, there have been few studies focusing on the optimal selection of Box–Cox transformation parameters in demographic forecasting. A simple method is proposed for selecting the optimal Box–Cox transformation parameter, along with an algorithm based on an in-sample forecast error measure. Illustrated by Australian age-specific fertility, the out-of-sample accuracy of a forecasting method can be improved with the selected Box–Cox transformation parameter. Furthermore, the log transformation is not adequate for modelling and forecasting age-specific fertility. The Box–Cox transformation parameter should be embedded in statistical analysis of age-specific demographic data, in order to fully capture forecast uncertainties.

Keywords Age-specific fertility rates · Data transformation · Principal component analysis · Mean absolute forecast error · Interval score

Introduction

In the demographic literature, forecasting methods for age-specific fertility can be generally grouped into parametric, semiparametric and nonparametric models.

Electronic supplementary material The online version of this article (doi:10.1007/s12546-014-9138-0) contains supplementary material, which is available to authorized users.

H. L. Shang (✉)
Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University, Canberra, ACT 0200, Australia
e-mail: hanlin.shang@anu.edu.au

Parametric models used in forecasting include the beta, gamma, double exponential and Hadwiger functions (Congdon 1990, 1993; Keilman and Pham 2000; Knudsen et al. 1993; Thompson et al. 1989), while semiparametric models include the Coale–Trussell and Relational Gompertz models (Booth 1984; Brass 1981; Coale and Trussell 1974; Murphy 1982; Zeng et al. 2000). The use of these models is variously limited by parameter un-interpretability, over-parameterization and the need for vector autoregression; structural change also limits their utility, especially where vector autoregression is involved (Booth 2006). To address this problem, nonparametric methods use a dimension-reduction technique, such as principal components analysis, to linearly transform age-specific fertility rates to extract a series of time-varying indices to be forecast (see Bell 1992; Bozik and Bell 1987; Hyndman and Ullah 2007; Lee 1993).

The Box–Cox transformation can sometimes yield noticeable improvements in model simplicity, variance homogeneity and precision of estimation. Despite the rapid development in demographic forecasting models, there have been few studies focusing on the optimal selection of the Box–Cox transformation parameter, with the noticeable exception of Hyndman and Booth (2008). As noted in early work by Box and Cox (1964) and Box (1988), the careful selection of a data transformation is often treated as a prerequisite before any serious modelling takes place.

An example of data transformation is the log transformation for modelling and forecasting age-specific mortality. Such a transformation allows researchers to visualize and model patterns associated with the 'accident bump' and to exploit near-linearities in the log mortality rates for the ages 40–80 years. The log transformation is a special case of the Box–Cox transformation, which can be defined as

$$z_{t,i} = \begin{cases} \frac{1}{\lambda}\left[\left(f_{t,i}\right)^{\lambda} - 1\right], & \lambda \neq 0 \\ \ln\left(f_{t,i}\right), & \lambda = 0 \end{cases}$$

where $f_{t,i} > 0$ denotes the observed age-specific data at age i in year t, whereas $z_{t,i}$ denotes the transformed data, and λ is the transformation parameter. For instance, when $\lambda = 1$, the transformation is essentially the identity, and the logarithm when $\lambda = 0$. In this work, we restrict it to lie in the unit interval (see also Hyndman and Booth 2008).

I propose a simple and instructive way to select the optimal Box–Cox transformation parameter based on an in-sample forecast error measure, and to demonstrate this idea in the context of modelling and forecasting age-specific fertility. The effect of the Box–Cox transformation on fertility is mainly manifested by a different shape of age profile. With the optimal transformation parameter, the age profile of the transformed data may reveal age patterns that are not obvious in the raw data.

This paper is organized as follows: first I present the Australian age-specific fertility from 1921 to 2006, followed by methodology and optimization algorithm. Results are then shown and the final section concludes with some thoughts on how the method developed here might be further extended.

Data and design

Data

We consider annual Australian age-specific fertility rates from 1921 to 2006. The dataset has been obtained from the Australian Bureau of Statistics (2008), and is also available in the rainbow package (Shang and Hyndman 2013) in R Core Team (2014). The data consist of annual fertility rates by single-year age of mother aged from 15 to 49 years. A graphical data display is given in Fig. 1. From the rainbow plot in Fig. 1a, we see the phenomenon of fertility postponement in the most recent years. From the contour plot in Fig. 1b, we see the increases in fertility between ages 20 and 30 from 1940 to 1980; this reflects the baby boom period.

As a demonstration, Fig. 2 presents the Box–Cox transformed fertility rates for years 1921 and 2006. With different values of λ, the age patterns change accordingly. The goal is to select the optimal λ that improves model estimation and prediction accuracy for a chosen model.

Study design

Since the optimal Box–Cox transformation parameter is selected on the basis of an in-sample forecast error measure, we divide the data into a training sample, a validation sample and a testing sample. Customarily, the testing sample consists of the last 20 % of the data, which are used to examine the out-of-sample forecast accuracy with the selected Box–Cox transformation parameter. The validation sample, which has the same number of data as the testing sample, is used to select the optimal Box–Cox transformation parameter. As in the case of Australian fertility rates, the training sample is from 1921 to 1972, the validation sample is from 1973 to 1989, and the testing sample is from 1990 to 2006.

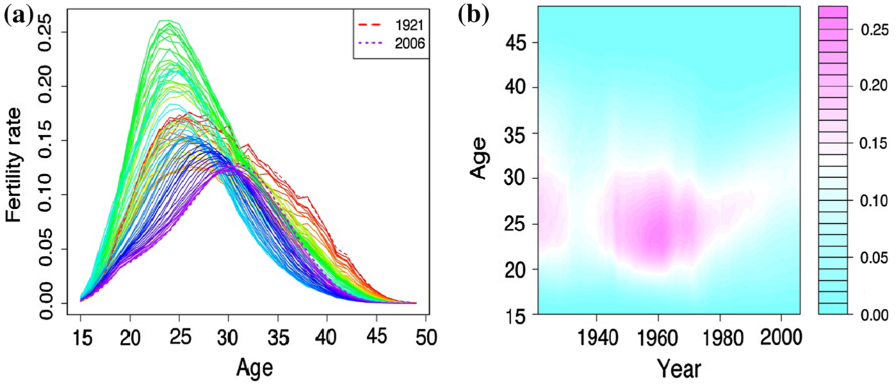
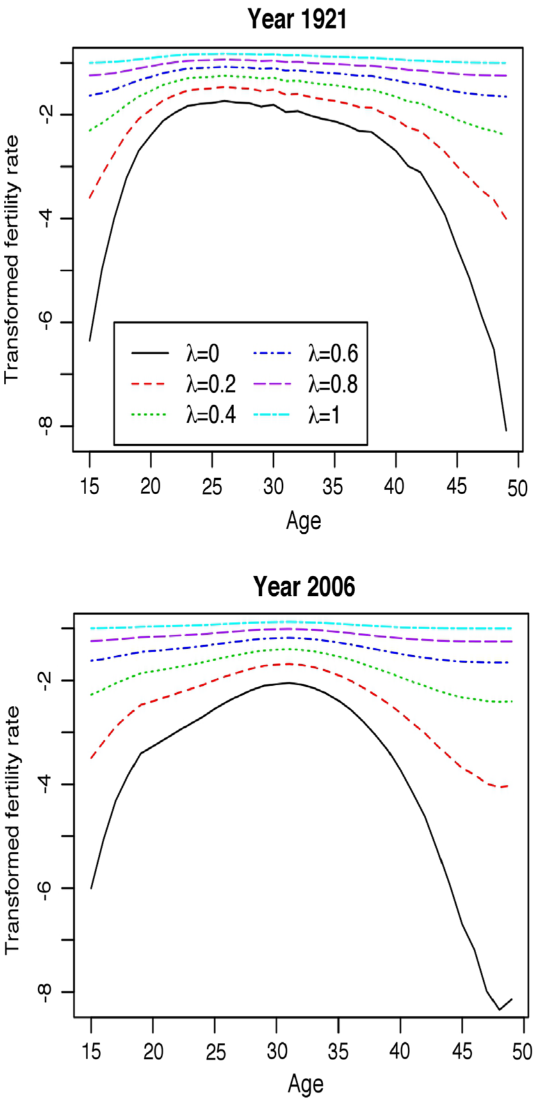


Fig. 1 Observed age-specific fertility rates for Australia from 1921 to 2006. **a** Rainbow plot of fertility rates, **b** filled contour plot of fertility rates. *Note* that the *dashed line* represents the data in 1921, while the *dotted line* represents the data in 2006. *Source*: Australian Bureau of Statistics (Cat. No. 3105.0.65.001, Table 38)

Fig. 2 Box–Cox transformed fertility rates with different values of λ in years 1921 and 2006, as two examples



There are various ways to measure the forecast accuracy. Following the early work by Shang et al. (2011), we use mean absolute forecast error (MAFE) for measuring point forecast accuracy. This is given by

$$\mathrm{MAFE}_h = \frac{1}{(18-h)\times 35}\sum_{s=1}^{18-h}\sum_{i=15}^{49}|f_{i,s}-\hat{f}_{i,s}|, \quad h = 1, \ldots, 17.$$

The MAFE is the average of absolute error across ages and years in the forecasting period; it measures forecast precision regardless of sign and is not sensitive to large relative errors of small rates. Since the back-transformed forecasts are median forecasts on the original scale, this makes them suitable for evaluation using MAFE.

In order to evaluate the interval forecast accuracy, we use the interval score of Gneiting and Raftery (2007) and Gneiting and Katzfuss (2014). For each year in the forecasting period, the 1-step-ahead to 17-step-ahead prediction intervals were calculated at the $(1-\alpha)\times 100\,\%$ nominal coverage probability. We consider the common case of the symmetric $(1-\alpha)\times 100\,\%$ prediction interval, with lower and upper bounds that are predictive quantiles at $\alpha/2$ and $1-\alpha/2$, denoted by l and u. As defined by Gneiting and Raftery (2007), a scoring rule is given by the associated interval forecast $S_\alpha\ (l,\ u;\ i)$. This can be expressed as

$$S_\alpha(l,u;i) = (u-l) + \frac{2}{\alpha}[(l-i)I\{i<l\} + (i-u)I\{i>u\}],$$

where $I\{\cdot\}$ represents the binary indicator function which takes the value of 1 when the condition is met, and α denotes the level of significance. In this paper, $\alpha = 0.2$ since I construct an 80 % prediction interval. The optimal score is achieved when i lies between l and u, and the distance between l and u is minimal. The interval score can be interpreted as follows: a forecaster is rewarded for narrow width of a prediction interval, if and only if the true observation lies within the prediction interval. The smaller the interval score is, the better the method is for producing interval forecasts.

For different ages and years in the forecasting period, the averaged interval score is defined by

$$S_\alpha^{\text{ave}}(l,u;i) = \frac{1}{(18-h)\times 35}\sum_{s=1}^{18-h}\sum_{i=15}^{49} S_{\alpha,s}(l,u;i), \quad h = 1,\ldots,17.$$

Method

Many methods have been proposed for modelling age-specific fertility (see Booth 2006). To demonstrate our main idea, we model the observed period age-specific fertility, using the well known Lee–Carter model (Lee and Carter 1992). Instead of retaining only the first component, we retain more than one principal component (see also Cairns et al. 2006). The modified Lee–Carter model can be defined by

$$z_{t,i} = \mu_i + \sum_{k=1}^{K}\beta_{t,k}\Phi_{k,i} + \varepsilon_{t,i}, \quad 1\le t\le n, \quad 1\le i\le p,$$

where n denotes the last year in the training sample and p denotes the last age, μ_i represents the mean estimated by $\frac{1}{n}\sum_{t=1}^{n} z_{t,i}$, $\{\beta_{1,k},\ldots\beta_{n,k}\}$ represents the kth estimated principal component scores, $\{\Phi_{k,1},\ \ldots,\ \Phi_{k,p}\}$ represents the kth estimated principal component which can be obtained from singular value decomposition applied to the training sample, $\varepsilon_{t,i}$ represents the independent and identically distributed Gaussian white noise, and K represents the number of retained principal components and the value of K can be determined by a ratio-based estimator (see Lam et al. 2011).

Point and interval forecasts

Conditional on the estimated mean $\hat{\mu}_i$ and the estimated principal components $\left(\hat{\Phi}_{1,i}, \ldots, \hat{\Phi}_{K,i}\right)$, the point forecasts are given by

$$\hat{z}_{n+h|n,i} = \hat{\mu}_i + \sum_{k=1}^{K} \hat{\beta}_{n+h|n,k} \hat{\Phi}_{k,i},$$

where $\hat{\beta}_{n+h|n,k}$ represents the h-step-ahead point forecast of the kth principal component scores. These forecasts can be obtained from applying a univariate time-series model, such as an autoregressive integrated moving average (ARIMA) model. We use the *auto.arima* algorithm of Hyndman and Khandakar (2008) to select the optimal orders of an ARIMA model on the basis of an information criterion, such as the corrected Akaike information criterion (Hurvich and Tsai 1989) considered in this paper.

Similarly, conditional on the estimated mean and estimated principal components, the total variance can be approximated by

$$\text{Var}\left[\hat{z}_{n+h|n,i}\right] \approx \sum_{k=1}^{K} \hat{\mu}_{n+h|n,k} \hat{\Phi}^2_{k,i} + \hat{v}_{n+h,i},$$

where $\hat{\mu}_{n+h|n,k}$ denotes the estimated variance of the sample principal component scores; $\hat{\Phi}^2_{k,i}$ denotes the square of the fixed principal components; and $\hat{v}_{n+h,i}$ denotes the estimated variance of the model residual (see also Shang et al. 2011). The 80 % prediction interval of the transformed data can be obtained on the basis of the estimated total variance and a normality assumption.

Having obtained the point and interval forecasts for the transformed data, we then back-transform these forecasts to the original scale through inverse Box–Cox transformation. This can be expressed as

$$\hat{f}_{n+h|n,i} = \begin{cases} \left(\lambda \hat{z}_{n+h|n,i} + 1\right)^{\frac{1}{\lambda}}, & \textit{if } \lambda \neq 0 \\ \exp\left(\hat{z}_{n+h|n,i}\right), & \textit{if } \lambda = 0 \end{cases}$$

Application to age-specific fertility

In Fig. 3, we present principal components and their associated scores for the Australian fertility data from 1921 to 1989. Based on these data, the forecasts of fertility rates from 1990 to 2006 are obtained. Although the optimal K is selected by the ratio-based estimator, only the first two components are displayed for ease of presentation. In the top panel, we show the age profile. In the middle panel, we display the time trend of the principal component scores. In particular, the point forecasts of the scores are shown in solid line, whereas the dark and light grey

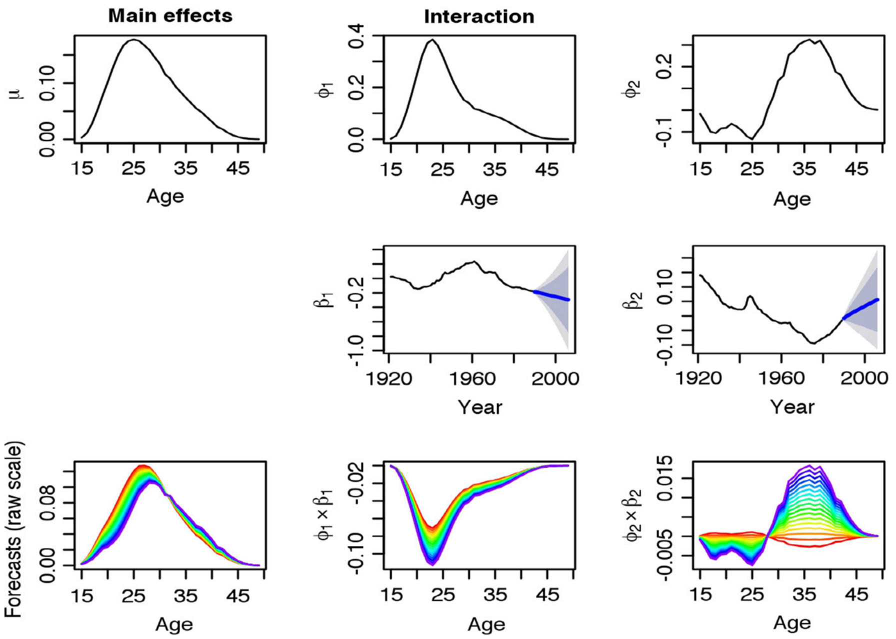


Fig. 3 Principal component decomposition for the Australian fertility data from 1921 to 1989, from which the forecasts are obtained from 1990 to 2006

regions represent the 80 % and 95 % pointwise prediction intervals, respectively. In the bottom panel, the forecasts of fertility are obtained by multiplying the fixed principal components by the forecast principal component scores before adding the main effect.

The first principal component models the fertility rates at early ages, whereas the second principal component models the fertility rates at later ages. From the forecast first principal component scores, it is clear that the fertility trend at early ages is likely to decline. From the forecast second principal component scores, it is evident that the fertility trend at later ages is likely to increase.

Results

Selection of the optimal parameter

The previous section presents one method for modelling and forecasting age-specific fertility, but the main contribution is to present a method to select the optimal transformation parameter based on in-sample forecast accuracy. To investigate the in-sample forecast accuracy, we implement the rolling origin approach. Using the initial training sample in the Australian age-specific fertility, we produce 1- to 17-step-ahead point and interval forecasts. Then, we increase the sample size by 1 year, re-estimate the model and produce 1- to 16-step-ahead

Table 1 The estimated optimal Box–Cox transformation parameters and out-of-sample point and interval forecast accuracy for different horizons

h	Point forecast accuracy			Interval forecast accuracy	
	$\text{MAFE}_{\lambda=0.46}$	$\text{MAFE}_{\lambda=0}$	$\text{MAFE}_{\lambda=0.4}$	$\text{Score}_{\lambda=0.46}$	$\text{Score}_{\lambda=0}$
1	**0.00117**	0.00235	0.00120	**0.00543**	0.00682
2	**0.00152**	0.00304	0.00155	**0.00732**	0.00982
3	**0.00219**	0.00388	0.00225	**0.00936**	0.01332
4	**0.00285**	0.00487	0.00289	**0.01174**	0.01696
5	**0.00352**	0.00590	0.00360	**0.01412**	0.01905
6	**0.00414**	0.00721	0.00432	**0.01651**	0.02135
7	**0.00487**	0.00853	0.00508	**0.01942**	0.02568
8	**0.00564**	0.00964	0.00591	**0.02134**	0.02909
9	**0.00635**	0.01067	0.00662	**0.02400**	0.03324
10	**0.00697**	0.01180	0.00735	**0.02610**	0.03743
11	**0.00758**	0.01291	0.00780	**0.02865**	0.04313
12	**0.00819**	0.01400	0.00844	**0.03097**	0.04745
13	**0.00894**	0.01499	0.00938	**0.03314**	0.05058
14	**0.00962**	0.01668	0.01013	**0.03584**	0.06074
15	**0.01032**	0.01782	0.01070	**0.03912**	0.06641
16	**0.01068**	0.01871	0.01111	**0.04118**	0.07244
17	**0.00992**	0.01830	0.01179	**0.04337**	0.07312
Mean	**0.00614**	0.01067	0.00648	**0.02398**	0.03686
Median	**0.00635**	0.01067	0.00662	**0.02400**	0.03324

The minimum MAFE and interval score are highlighted in bold for each horizon and the summary statistics

forecasts. This process is iterated until the training sample reaches the last year of the validation sample. This would produce 17 one-step-ahead forecasts, 16 two-step-ahead forecasts, up to one 17-step-ahead forecast. We use these forecasts to evaluate the out-of-sample forecast accuracy. For a range of forecast horizons, we calculate its forecast accuracy based on an error measure, such as MAFE or interval score given in (1), over different ages and years in the validation sample. The optimal Box–Cox transformation parameter is the one that minimizes the median of a forecast error measure over a range of forecast horizons. Computationally, the optimization can be achieved by using the *optimize* function in R.

In Table 1, we present the selected Box–Cox transformation parameter λ, based on the in-sample MAFE and interval score. For the purpose of comparison, we also consider the log transformation which is commonly used in modelling age-specific mortality. From the averaged MAFE and averaged interval score across 17 horizons, we found that with the selected Box–Cox transformation parameter, the out-of-sample point and interval forecast errors can be reduced in comparison with the log transformation for each forecast horizon.

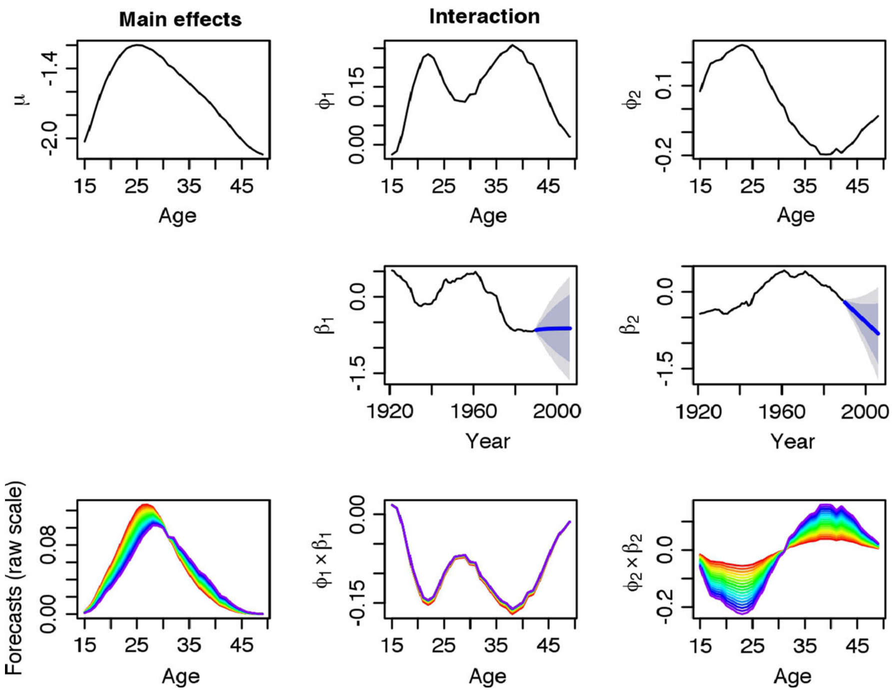


Fig. 4 Principal component decomposition for the transformed Australian fertility data from 1921 to 1989, from which the forecasts are obtained from 1990 to 2006

Note that Table 1 is consistent with the results of Hyndman and Booth (2008) who found that the best point forecast accuracy (for one-step-ahead forecasts) had $\lambda = 0.4$. In comparison to $\lambda = 0.4$, we found that our selected $\lambda = 0.46$ gives better accuracy for each horizon, but their differences in point forecast accuracy on the testing sample are marginal.

Application to age-specific fertility

Before fitting a modified Lee–Carter model, the raw data are transformed by the Box–Cox transformation, which may introduce a small bias, but can potentially reduce variance. As a result, this may improve estimation and forecast accuracy. In terms of its effect on forecasts of fertility, Fig. 4 displays the principal component decomposition for the Box–Cox transformed data with $\lambda = 0.46$. From the bottom left plot, it is evident that the forecast fertility rates have a similar shape to the ones using the raw data. However, the age patterns are very different in shape from the untransformed ones, such as the bimodality shown in the first principal component. From the forecast first principal component scores, such bimodality is likely to continue with increasing forecast uncertainties as the horizon increases. The second principal component shows the contrast between ages around 25 and 40. From the forecast second principal component scores, such contrast is likely to decrease with increasing forecast uncertainties as the horizon increases.

Conclusion and future research

This paper presented a method and an algorithm for selecting the optimal Box–Cox transformation parameter. The contributions of this paper are twofold: first, it was found that the log transformation may not be adequate for modelling and forecasting age-specific fertility. Second, a way of selecting the optimal Box–Cox transformation parameter based on in-sample forecast accuracy was presented and showed that with the selected Box–Cox transformation parameter, the out-of-sample point and interval forecast accuracy can be improved. In addition, the optimal Box–Cox transformation parameter $\lambda = 0.46$ produces slightly smaller point forecast error in comparison to $\lambda = 0.4$ used in Hyndman and Booth (2008).

The proposed method and algorithm can be extended to select the optimal Box–Cox transformation parameter for modelling and forecasting age-specific migration. With the selected Box–Cox transformation parameter, the forecast uncertainties associated with age-specific components of population change are more likely to be fully captured. Finally, from a Bayesian viewpoint, it is also possible to embed the selection of the optimal Box-Cox transformation parameter into the modelling and forecasting. The R code for implementing the proposed algorithm for the Australian age-specific fertility is available at the online supplementary material.

Acknowledgments The author thanks two referees for their insightful comments and suggestions, which led to a much improved manuscript. The author thanks Bridget Browne for proof-reading an early version of this manuscript.

References

Australian Bureau of Statistics (ABS). (2008). *Australian historical statistics*. Cat. No. 3105.0.65.001. Canberra.

Bell, W. (1992). ARIMA and principal components models in forecasting age-specific fertility. In N. Keilman & H. Cruijsen (Eds.), *National population forecasting in industrialized countries* (pp. 177–200). Amsterdam: Swets & Zeitlinger.

Booth, H. (1984). Transforming Gompertz's function for fertility analysis: The development of a standard for the relational Gompertz function. *Population Studies, 38*(3), 495–506.

Booth, H. (2006). Demographic forecasting: 1980–2005 in review. *International Journal of Forecasting, 22*(3), 547–581.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformation (with discussion). *Technometrics, 30*(1), 1–40.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformation. *Journal of the Royal Statistical Society: Series B, 26*(2), 211–252.

Bozik, J., & Bell, W. (1987). Forecasting age specific fertility using principal components. In *Proceedings of the American Statistical Association* (pp. 396–401). *Social Statistics Section*, San Francisco, CA.

Brass, W. (1981). The use of the Gompertz relational model to estimate fertility. In *Proceedings of the international population conference* (pp. 345–362). Manila: IUSSP.

Cairns, A. J. G., Blake, D., & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance, 73*(4), 687–718.

Coale, A. J., & Trussell, T. J. (1974). Model fertility schedules: Variations in the age structure of childbearing in human populations. *Population Index, 40*(2), 185–258.

Congdon, P. (1990). Graduation of fertility schedules: An analysis of fertility patterns in London in the 1980s and an application to fertility forecasts. *Regional Studies, 24*(4), 311–326.

Congdon, P. (1993). Statistical graduation in local demographic analysis and projection. *Journal of the Royal Statistical Society, Series A, 156*(2), 237–270.

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

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

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

Hyndman, R. J., & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting, 24*(3), 323–342.

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

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

Keilman, N., & Pham, D. Q. (2000). Predictive intervals for age-specific fertility. *European Journal of Population, 16*(1), 41–66.

Knudsen, C., McNown, R., & Rogers, A. (1993). Forecasting fertility: An application of time series methods for parameterized model schedules. *Social Science Research, 22*(1), 1–23.

Lam, C., Yao, Q., & Bathia, N. (2011). Estimation of latent factors in high-dimensional time series. *Biometrika, 98*(4), 901–918.

Lee, R. D. (1993). Modeling and forecasting the time series of US fertility: Age distribution, range and ultimate level. *International Journal of Forecasting, 9*(2), 187–202.

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association, 87*(419), 659–671.

Murphy, M. J. (1982). *Gompertz and Gompertz relational models for forecasting fertility: An empirical exploration*. Working paper, Centre for Population Studies, London School of Hygiene and Tropical Medicine, London.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Shang, H. L., Booth, H., & Hyndman, R. J. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research, 25*, 173–214.

Shang, H. L., & Hyndman, R. J. (2013). *Rainbow: Rainbow plots, bagplots and boxplots for functional data*. R package version 3.2. http://cran.r-project.org/web/packages/rainbow.

Thompson, P. A., Bell, W. R., Long, J. F., & Miller, R. B. (1989). Multivariate time series projections of parameterized age-specific fertility rates. *Journal of the American Statistical Association, 84*(407), 689–699.

Zeng, Y., Wang, Z., Ma, Z., & Chen, C. (2000). A simple method for projecting or estimating α and β: An extension of the Brass relational Gompertz fertility model. *Population Research and Policy Review, 19*(6), 525–549.