International Journal of Forecasting 32 (2016) 726–735

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Household forecasting: Preservation of age patterns

Nico Keilman

Department of Economics, University of Oslo, Norway

ARTICLE INFO

Keywords:
Household dynamics
Lee–Carter model
Brass relational method
Random walk with drift
The Netherlands
Age profiles

ABSTRACT

We formulate a time series model of household dynamics for different age groups. We model the shares of the population who are in certain household positions (living alone, living with a partner, etc.). These household positions have very pronounced age patterns. The age profiles change slowly over time, due to changes in the home leaving behaviour of young adults, differences in survival rates of men and women, etc. When forecasting household positions to 2040, we want to preserve the characteristics of the age profiles. We test the Lee–Carter model and the Brass relational method using household data for the Netherlands for the period 1996–2010. Annual shares of the population by household position, age, and sex are modeled as random walks with adrift (RWD). While the Brass model has its limitations, it performs better than the Lee–Carter model in our application. The predicted age patterns based on the Brass model look more reasonable, because the Brass model is a two-parameter model, while the Lee–Carter model contains only one parameter. Also, the model parameters and standard errors of the Brass model are easier to estimate than those of the Lee–Carter model.

1. Motivation

Fig. 1 shows, for the case of The Netherlands in 2011, the proportions of women who live with parents, alone, in a consensual union, or with a marital spouse, broken down by five-year age groups. Most adolescents live with their parents. Those who have left home most often live alone or in a consensual union, up to ages around 30. After that, living with a spouse becomes the dominant position, until ages around 70. Some women become lone mothers, due to separation or divorce. Next, increasing numbers lose their husbands because the husband is a few years older (aggravated by the higher mortality of men), and many elderly women live alone, or together with one or more children. Of women over 95, more than half live in an institution (not shown in the graph).

Age profiles of the type shown in Fig. 1, and their development over time, help us to understand household dynamics. This in turn, when combined with forecasts of future age structures, facilitates demographers in projecting the number of households of various types into the future. Combining population forecasts with future values of carefully selected sets of household parameters is a well-

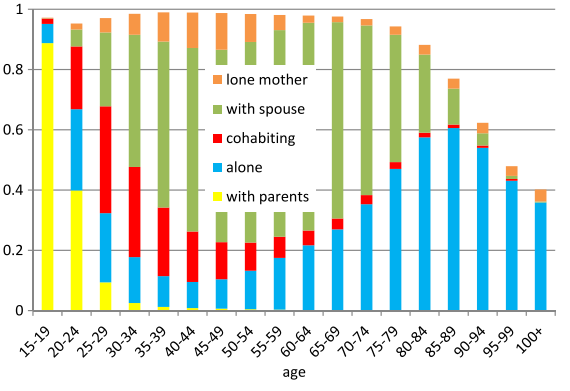

Fig. 1. Proportions of women living in various private household positions, by age; The Netherlands, 2011.
Source: Census data from Eurostat.

E-mail address: nico.keilman@econ.uio.no.

http://dx.doi.org/10.1016/j.ijforecast.2015.10.007

established method of computing household forecasts; see the extensive review by Holmans (2012). In many countries, the life expectancy of men is increasing faster than that of women. What does that imply for the numbers of elderly men and women who live alone? Business cycles and youth unemployment have effects on the home-leaving behaviour of young adults. Formal marriages have become less important in many Western countries since the 1970s, but did consensual unions fill the gap fully or only partially? These and related issues show that it is important to describe and understand the age profiles of various household parameters, when computing household forecasts to be used by policy makers in such diverse fields as housing, social security, consumption, and energy consumption, to name only a few.

Ideally, household forecasts should be based on well-established theories of the household behaviours of individuals. Many scholars have tried to develop social, economic and cultural theories to explain why households change over time. The reasons for such changes include a reduced adherence to strict norms; less religiosity and an increase in individual freedom on ethical issues; female education, which has led to women having greater economic independence, and also facilitates divorce; more assertiveness in favour of symmetrical gender roles; the contribution of women to the labour market; increased economic aspirations; and residential autonomy (Lesthaeghe, 1995; Van de Kaa, 1987; Verdon, 1998). In addition, there are also demographic factors, such as falling levels of fertility, and differences in longevity between men and women. However, none of these theories have resulted in formalized models of household behaviour that are general enough and have sufficient explanatory power to be used for forecasting. Two decades ago, Burch (1995) noted that methods for modelling family and household dynamics had made considerable progress, but that theory had lagged behind considerably. The situation is not much better today, which may reflect the complexity of the subject matter. Thus, as a second best to predicting households based on general behavioural theories, we look for regularities in the observed data, try to understand the trends, and extrapolate them into the future by means of formal time series models. Sometimes the forecaster has very little data, perhaps only one year's worth, upon which the forecast can be based. In that case, a commonly-used approach is simply to keep the parameters of interest constant over the forecast period. One example is the multi-state approach to modelling household dynamics (Van Imhoff & Keilman, 1991), in which the transition probabilities that describe changes among household positions for individuals are kept constant. In the current paper, however, we are able to use time series data over a longer period. This allows us to take possible time trends in the parameters into account explicitly. In addition (though we do not use this here), a time series approach also allows one to make stochastic predictions, and hence to take the prediction uncertainty into account.

The aim of this paper is to show how time series data for the age profiles of men and women in various household positions can be modelled. Using data from The Netherlands for the period 1996–2010, we model the vector of age-specific shares for a certain household position as a random walk with drift (RWD). In other words, we assume that the year-on-year step for the shares consists of a certain fixed term (the drift) plus a normally distributed error term which has a zero expectation. The result is a share with an increasing variance around a linear trend. A random walk with drift model is one type of time series model that has been applied to demography (e.g., Alho & Spencer, 2005). The book by Box and Jenkins (1976) is a standard reference for this and other types of time series models. If one were to model the share of, say, women who live alone as a RWD for each age separately, one would run the risk that the drift terms may be very different for different ages; see Christiansen and Keilman (2013). This would distort the age pattern for these women. In order to retain the age patterns for the household shares, we have selected two methods that were developed originally for mortality analysis, namely the Lee–Carter model (henceforth abbreviated as LC) and the Brass relational method (Brass). We then use the estimated RWD models to extrapolate the shares thirty years ahead.

The contribution this paper makes is to show how data reduction techniques, stemming from mortality analysis, can be used to describe and project household dynamics. More specifically, we show that the Brass method has considerable advantages over LC for this particular data set: the resulting age patterns for men and women in various household positions look more realistic, and the model parameters are easier to estimate.

The remainder of the paper proceeds as follows. In Section 2 we begin by describing the household data, which stem from the population registers of the Netherlands, then outline the LC-model and the Brass approach. Section 3 presents the estimation results, while Section 4 discusses the extrapolated age profiles for future years. We finish the paper in Section 5 with a discussion and conclusions.

2. Data and methods

We are interested in modelling household shares. Write $V(j, x, s, t)$ for the number of people in household position $j = 1, 2, \ldots$ who are of age $x = 0, 1, \ldots$ and sex $s = 1$ or 2, at time $t = 0, 1, 2, \ldots$. Aggregating over position, we obtain the population who are of age x and sex s at time t as $W(x, s, t) = \Sigma_j V(j, x, s, t)$. The share of household position j is $\alpha(j, x, s, t) = V(j, x, s, t)/W(x, s, t) = \alpha_j(x, s, t)$.

2.1. Data

Coen van Duin of Statistics Netherlands kindly supplied us with data from the population registers on the household positions of men ($s = 1$) and women ($s = 2$) in the Netherlands, broken down by age group ($x = 1$ for ages 0–4, $x = 2$ for ages 5–9, …, $x = 20$ for age 95+), as of 1 January of the years 1996 ($t = 1$), 1997 ($t = 2$), …, 2010 ($t = 15$). Following earlier work (Christiansen & Keilman, 2013), we distinguish seven mutually exclusive household positions that an individual can occupy at any given point in time. These household positions are

particularly relevant for household analyses in the context of demographic behaviour, as well as in studies of social security and consumption. In other applications, such as studies of housing needs, one may prefer to distinguish individuals according to household sizes. The household positions that we use in the current analysis are (with household position codes in parentheses):

$j = 1$. Child living with parent(s) (CHLD).
$j = 2$. Living in one-person household (SIN0).
$j = 3$. Living in unmarried cohabitation, with or without children (COH).
$j = 4$. Living with marital spouse, with or without children (MAR).
$j = 5$. Living as lone parent (SINp).
$j = 6$. Other position in private household, for instance a member of a multiple-family household, living with non-family-related individuals, or homeless (OTHR).
$j = 7$. Living in an institution (INST).

These categories refer to living arrangements, not marital status. For example, the category MAR does not include all of those who are married, but only those who are currently living with their spouse. One example of a person who would belong to the group OTHR is someone living in a multiple-family household. Persons who live in households with no parent–child relationship, and who are not married or cohabiting with any of the other members of the household, also belong to this category. No age restrictions have been imposed on persons in a certain household position. In particular, children (CHLD) and lone parents (SINp) can be of any age. In practice, persons aged 85, say, with positions CHLD or SINp, will not be interpreted as such, but should be assigned to a different position, for instance to the group OTHR. Moreover, we have ignored the few persons who are aged younger than 15 in the following household positions: SIN0, COH, MAR, and SINp.

2.2. Method—generalities

Before modelling the random evolution of the shares, a logit transformation was applied. We have opted for a hierarchy of household positions using a variant of continuing fractions. This led to there being six types of fraction requiring modelling (all specific to age, sex and time), starting from the shares $\alpha_j(x, s, t)$ defined in the first paragraph of Section 2. The following fractions were defined, given age, sex, and time (with household position codes in parentheses):

1. The share of CHLD (CHLD);
2. The relative share of COH and MAR out of the total share of one minus the share of CHLD (COHMAR);
3. The relative share of MAR out of the share of COH and MAR (MAR);
4. The relative share of SIN0 and INST out of the total share of SIN0, SINp, OTHR, and INST (SIN0INST);
5. The relative share of SIN0 out of the share of SIN0 and INST (SIN0);
6. The relative share of SINp out of the total share of SINp and OTHR (SINp).

Because of the hierarchy, the predicted shares in the logit scale at a higher level are independent of those at a lower level. The particular sequence 1–6 above is based upon the idea that important shares (numerically, behaviourally) have to be modelled first, and those that are less important can come last. Hence, persons who live together with a partner (points 2 and 3 above), or alone (points 4 and 5) are given priority. For elderly persons, the position INST is often difficult to separate from positions in non-institutional households, due to unclear registration rules for persons who *de facto* live in an institution (Christiansen & Keilman, 2013). Thus, they are dealt with initially as one group (point 4), taking into account the fact that positions COH and MAR have been covered already. Children are singled out from the beginning, because their shares are kept constant over time. The age pattern for this household position shows very little variation: for ages under 15, the shares are almost 100% (some children live in multi-family households and hence have household position OTHR, a few live in an institution). The shares then fall rapidly for ages 15–19 and 20–24, and are close to zero for ages beyond 25. Hence, systematic changes over time in the age patterns are difficult to identify. Finally, note that we have selected the household position OTHR as a remainder, which is in agreement with the nature of this position as we have defined it.

Temporarily suppressing the indices for age, sex, and time, the logit transforms of the fractions 2–6 above are

$$\xi_2 = \text{logit}((\alpha_3 + \alpha_4)/(1 - \alpha_1))$$
$$\xi_3 = \text{logit}(\alpha_4/(\alpha_3 + \alpha_4))$$
$$\xi_4 = \text{logit}((\alpha_2 + \alpha_7)/(\alpha_2 + \alpha_5 + \alpha_6 + \alpha_7))$$
$$\xi_5 = \text{logit}((\alpha_2)/(\alpha_2 + \alpha_7))$$
$$\xi_6 = \text{logit}(\alpha_5/(\alpha_5 + \alpha_6)),$$

where shares α_j are as defined at the beginning of Section 2. Thus, five series (given age and sex) were constructed.

There are many equivalent expressions for the back-transformation, i.e., for the α_j written as functions of the ξ_m. One of these is the following set:

$$\alpha_2 = (1 - \alpha_1) \cdot \exp(\xi_4) \cdot \exp(\xi_5)/ \{(1 + \exp(\xi_2))(1 + \exp(\xi_4))(1 + \exp(\xi_5))\}$$
$$\alpha_3 = (1 - \alpha_1) \cdot \exp(\xi_2)/\{(1 + \exp(\xi_2))(1 + \exp(\xi_3))\}$$
$$\alpha_4 = \alpha_3 \cdot \exp(\xi_3)$$
$$\alpha_6 = (1 - \alpha_1 - \alpha_3 - \alpha_4)/\{(1 + \exp(\xi_4))(1 + \exp(\xi_6))\}$$
$$\alpha_5 = \alpha_6 \cdot \exp(\xi_6)$$
$$\alpha_7 = \alpha_6 \cdot \exp(\xi_4)(1 + \exp(\xi_6))/(1 + \exp(\xi_5)).$$

By assumption, the share of children (α_1) is independent of ξ_m ($m = 2, 3, \ldots, 6$).

There are many possible modelling strategies for the age patterns of these household shares. One is to assume that each share in either the original scale ($\alpha_j(x, s, t)$; $j = 2, 3, \ldots, 7$) or the logit-scale ($\xi_m(x, s, t)$; $m = 2, 3, 4, 5, 6$) can be written as a mathematical function of age, with a functional form that may depend on m and s, and with parameters that may be time-dependent. Functions that have been used in demography include the Gompertz, Makeham, and Helligman-Pollard functions for mortality, the

Beta, Gamma and Hadwiger functions for fertility, Coale and McNeil's double-exponential function for nuptiality, and many others. Booth (2006) provides an extensive review of the most important functions for modelling the age patterns of vital events. She also reviews a more flexible approach, namely the so-called relational method. Here, one chooses a smooth age pattern as a standard, and then specifies a simple model that describes how the current age pattern differs from the standard. The first such model was developed by Brass (1971) in the context of mortality. Booth (1984) and Zeng et al. (2000) also used this approach to model the age pattern of fertility. Other applications of the relational approach include the Coale–Trussell model for fertility (Coale & Trussell, 1974) and De Beer's TOPALS approach, which has been applied to age patterns of both fertility and mortality (De Beer, 2011, 2012). The relational method is more flexible than the approach based on mathematical functions, because it requires fewer parameters (unless one views the standard age pattern as a series of parameters as well).

Of the many methods that have been developed, we selected the Lee–Carter model and the Brass relational method. To the best of our knowledge, neither of these has been applied to the modelling of household dynamics (however, see Zeng et al., 2000, for an application to marital status and home-leaving). Both belong to the tradition of relational approaches. The Lee–Carter model was selected because it has become very popular in recent years, and has been applied to a wide variety of situations and data sets, but not to household dynamics. The Brass method was selected because of its simplicity.

2.3. The Lee–Carter model

Lee and Carter (1992) originally developed their model for describing and predicting age-specific mortality. The LC model assumes that the logarithm of the mortality rate $m(x, t)$ for age x during year $t (x = 1, 2, \ldots, X;\ t = 1, 2, \ldots, T)$ can be written as

$$\ln(m(x, t)) = a(x) + b(x) \cdot k(t) + e(x, t). \tag{1}$$

Eq. (1) tells us that the rate $m(x, t)$ in logarithmic form is a function of a general age profile $a(x)$ and a time trend $k(t)$. The time trend is not the same for all ages, but is modified with an age profile $b(x)$. This indicates how the different age groups react to mortality change. The error term $e(x, t)$, with expectations equal to zero and a variance that is independent of both x and t, captures factors that are not included in the model.

Without further constraints, the parameters $b(x)$ and $k(t)$ are not unique. For instance, when $b(x)$ and $k(t)$ satisfy Eq. (1), then $c \cdot b(x)$ and $k(t)/c$ also satisfy Eq. (1) for any non-zero constant c. In order to obtain unique parameter estimates, one usually adds identifying constraints to Eq. (1). Many authors follow Lee and Carter and assume $\Sigma_x b(x) = 1$ and $\Sigma_t k(t) = 0$. The restriction on $k(t)$ implies that $a(x)$ can be estimated as the average log-rate $\Sigma_t \ln(m(x, t))/T$, where the average is taken over time. This shows that one can interpret $a(x)$ as a standard age schedule in the sense of relational methods. Since there are no regressors in the right hand side term of Eq. (1), ordinary regression cannot be used for estimating the parameters $b(x)$ and $k(t)$. Instead, Lee and Carter, and many authors since, estimated the parameters by singular value decomposition (SVD) of the matrix formed by subtracting $a(x)$-estimates from $\ln(m(x, t))$. Briefly, the singular value decomposition of an $m \times n$ matrix M is a factorization

$$M = U \cdot \Sigma \cdot V^T, \tag{2}$$

where U is an $m \times m$ orthogonal matrix, Σ is an $m \times n$ rectangular diagonal matrix (only the entries $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{ss}$ of Σ are non-zero ($s = \min(m, n)$) which contains the singular values of M, and V^T is the transpose of an $n \times n$ orthogonal matrix V. The first column of U times σ_{11} times the first row of V^T has the best rank-1 approximation to the input matrix M.

The LC-model has turned out to be particularly attractive for many mortality applications because the estimated time trend is roughly linear. Lee and Carter modelled the estimated time series $k(t)$ as a RWD, and used the extrapolated $k(t)$ values to predict age-specific mortality rates for future years.

Following these ideas, we assumed that the logit-transformed fractions $\xi(x, t)$ for each case m as defined in Section 2.2 ($m = 2, 3, 4, 5, 6$) and for men and women can be written as

$$\xi(x, t) = a(x) + b(x) \cdot k(t) + e(x, t). \tag{3}$$

Here, we have suppressed the indices for m and sex. The model was estimated by SVD, with identifying constraints for $b(x)$ and $k(t)$ and assumptions for $e(x, t)$, as stated above. Instead of the usual RWD model, we initially assumed a slightly more general model for the time index $k(t)$. The assumption was that the time index would follow an autoregressive model of order 1 (AR1), including a constant term. In other words,

$$k(t + 1) = D + \rho \cdot k(t) + d(t), \tag{4}$$

where ρ is the autoregressive parameter, and $d(t)$ is an error term with zero expectation, zero autocorrelation, and constant variance. In many cases, the estimate of ρ turned out to be very close to one. If one assumes that ρ equals one, Eq. (4) reduces to a RWD-process, i.e., $k(t + 1) = D + k(t) + d(t)$, where D is the drift. Under such a model, the time-increment $\Delta\xi(x, t) = \xi(x, t + 1) - \xi(x, t)$ for a given age x can be written as

$$\Delta\xi(x, t) = b(x) \cdot \{D + d(t)\} + \Delta e(x, t). \tag{5}$$

Eq. (5) can be interpreted as a random walk with drift (Girosi & King, 2008). The drift equals $b(x) \cdot D$, while the random part consists of innovations $b(x) \cdot d(t) + \Delta e(x, t)$.

Starting from a known value $\xi(x, T)$ in year T, a future value $\xi(x, T + h)$ h years ahead ($h = 1, 2, 3, \ldots$) is

$$\xi(x, T + h) = a(x) + b(x) \cdot k(T + h) + e(x, T + h). \tag{6}$$

The RWD-assumption for the time index implies that $k(T + h) = k(T) + h \cdot D + d(T) + \cdots d(T + h - 1)$. Inserting this into Eq. (6) gives

$$\xi(x, T + h) = a(x) + b(x) \cdot \{k(T) + h \cdot D + d(T) + \cdots d(T + h - 1)\} + e(x, T + h). \tag{7}$$

A forecast for h years ahead can be computed as

$$E[\xi(x, T+h)] = \hat{a}(x) + \hat{b}(x) \cdot \{\hat{k}(T) + h \cdot \hat{D}\}, \quad (8)$$

where $E[.]$ denotes expectations and $a(x)$, $b(x)$, $k(T)$, and D have been replaced with their estimated values.

2.4. The Brass relational method

We have used a Brass type of relational model for the transformed shares. The Brass relational model was originally intended for modelling age-specific survival from birth to age x, and can be written as

$$Y(x) = a + b \cdot Y^S(x) + e(x),$$

where $Y(x)$ is the logit-transformed probability of survival from birth to age x, while $Y^S(x)$ is some standard age pattern of survival, also in logit form. a and b are coefficients to be estimated from the data, and $e(x)$ is an error term. The model is linear in its parameters, and hence, one can estimate them using an ordinary least squares (OLS) regression. Changing the parameter a shifts the age pattern up or down, while b changes its slope. See e.g. Preston, Heuveline, and Guillot (2001) for a thorough discussion.

In a first stage, we used an OLS regression to estimate the Brass relational model applied to the age pattern of logit-transformed fractions $\xi_m(x, s, t)$ ($m = 2, 3, 4, 5, 6$) as defined above, for each year, and for men and women separately. For each m, the standard age pattern $\xi^S(x)$ was defined as the average value of $\xi(x, t)$, where the average was taken over all years t, for a given combination of age and sex. Hence, for each m, we obtained estimates of the parameters a and b that varied over time and between sexes. However, in most cases we noticed a gradual increase or decrease in the estimates of a and b over time. This suggested that a and b could be written as linear functions of time, i.e.

$$\xi(x, t) = (A + a \cdot t) + (B + b \cdot t) \cdot \xi^S(x) + e(x, t),$$

dropping the distinction by sex. In order to avoid spurious correlations, we detrended this model by taking first differences, and found

$$\Delta\xi(x, t) = a + b \cdot \xi^S(x) + d(x, t), \quad (9)$$

where $\Delta\xi(x, t) = \xi(x, t) - \xi(x, t-1)$, and $d(x, t) = \Delta e(x, t)$ is an error term.

Eq. (9) defines $\xi(x, t)$ as a random walk with drift (RWD). The drift $a + b \cdot \xi^S(x)$ consists of two parts: one part (a) is common for all ages, whereas the other ($b \cdot \xi^S(x)$) is an age-specific part. The term $\xi^S(x)$ preserves the age pattern in the random walk increments for each type of fraction m. The innovation variance is $\sigma^2 = \mathrm{Var}[d(x, t)]$. In a second stage, the results of which are reported below, we estimated the parameters a and b by OLS regression (across x), assuming an innovation variance that is independent of age and time.

Starting from a known value $\xi(x, T)$, a future value h years ahead ($h = 1, 2, \ldots$) is

$$\xi(x, T+h) = \xi(x, T) + h \cdot (a + b \cdot \xi^S(x)) + d(x, T+1) + \cdots + d(x, T+h).$$

Hence, a forecast h years ahead can be computed as

$$E[\xi(x, T+h)] = \xi(x, T) + h \cdot (\hat{a} + \hat{b} \cdot \xi^S(x)), \quad (10)$$

where a and b have been replaced with their estimated values.

Note the difference between Eq. (5), based on the Lee–Carter approach, and Eq. (9), with the Brass relational method as a starting point. Both can be interpreted as RWD-models. However, as was noted above, the drift of the Brass-RWD in Eq. (9) consists of one part that is common for all ages x, and another part that is age-specific. On the other hand, the drift of the LC-RWD in Eq. (5) is age-specific only.

3. Estimations

3.1. Lee–Carter

We fitted Eq. (3) to the logit-fractions ξ_2 to ξ_6 by means of SVD, by age group (15–19, 20–24, …, 90–94, 95+) and sex, for the years 1996–2010. Fig. 2 plots the fit for the first and last years in the period. The fits are excellent, without exception. This is not surprising, because in each case (for instance men, $m = 2$) there are 17 age groups and 15 years of observations, making a total of 255 observations. These are modelled by 17 (for $a(x)$) + 17 (for $b(x)$) + 15 (for $k(t)$) = 49 parameters, which makes an extremely high parameters-to-observations ratio of 0.19. For later reference, note that the age profiles for $m = 2$ (COHMAR), $m = 3$ (MAR), and $m = 4$ (SIN0INST) cross between 1996 and 2010. The trend was downwards at some ages, but upwards at others.

Fig. 3 plots annual estimates of the time index $k(t)$ as dots, with an assumed RWD-process for $k(t)$ as a straight line. The first row of the panel shows that living as a couple has become less frequent over the period, both for the combined household position of cohabiting and married (COHMAR; $m = 2$), and for the position married given that one lives with a partner (MAR; $m = 3$). Living alone or in an institution (SIN0INST; $m = 4$) has clearly become more frequent, particularly for men, which is driven by the increasing importance of living alone (SIN0; $m = 5$); cf. the second row of plots. There is no clear time pattern for lone fathers (SINp; third row), but lone mothers became more frequent from the beginning of the 21st century.

The random walk with drift process was fitted as a straight line between the first and last estimates. One important assumption for such a RWD process is that the innovations $d(t)$ are uncorrelated. However, there are a number of cases in which this assumption is not realistic. Strong autocorrelation is visible in the plots for $m = 2$ (COHMAR) for women, $m = 5$ (SIN0), and $m = 6$ (SINp) for women. In such cases, an autoregressive model seems to be more appropriate than a RWD model. In addition, the plot for the case of $m = 6$ (SINp) for men shows no systematic drift, meaning that it is unrealistic to assume a random walk process with a non-zero drift.

Table 1 reports parameter estimates for an autoregressive model of order 1 (AR1), and a random walk with drift. It turns out that the constant term of the AR1 model is

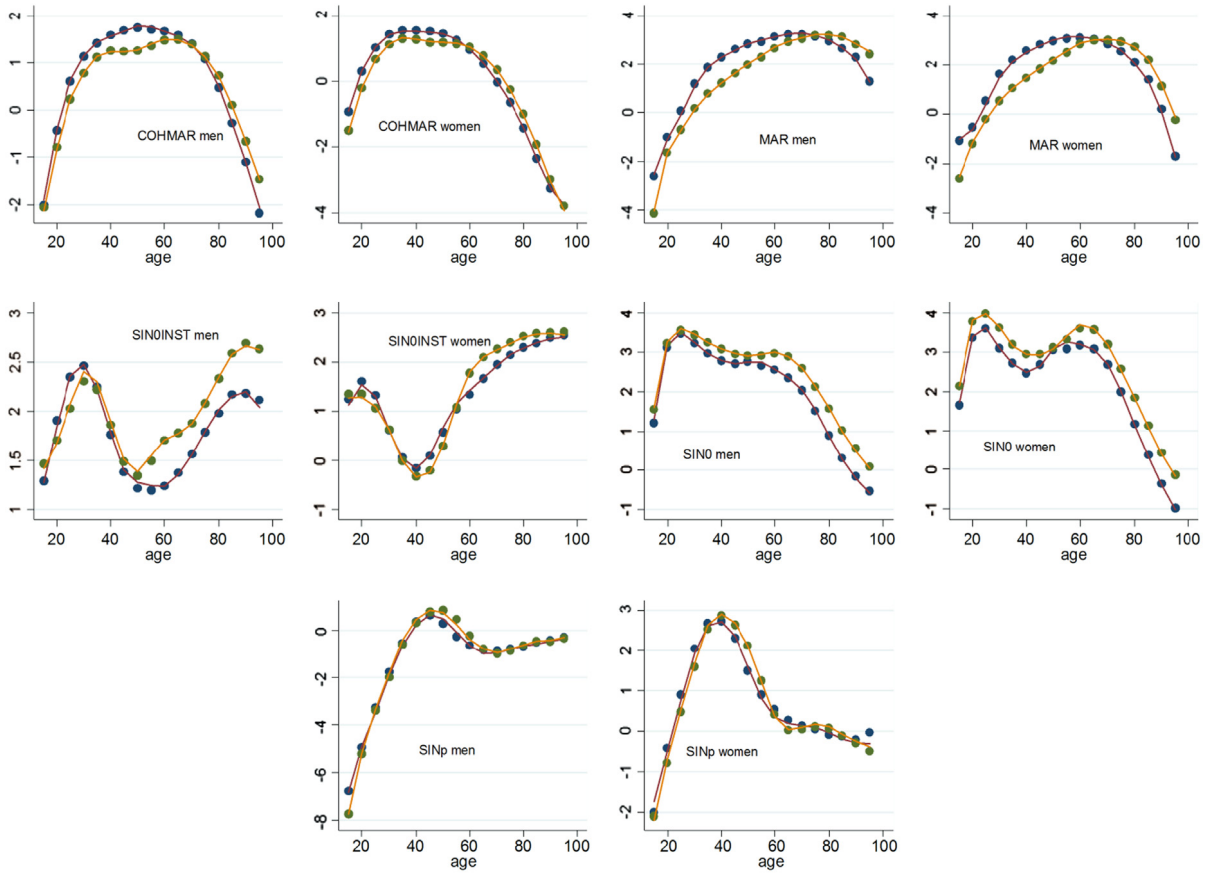

Fig. 2. Fit of logit of fractions to a Lee–Carter model for $m = 2$ (COHMAR) to $m = 6$ (SINp), men and women, 1996 (red line, blue dots) and 2010 (yellow line, green dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Parameter estimates for two time series models for the time index $k(t)$ of the LC model: a first-order autoregressive model (AR1) with constant term D and autoregressive parameter ρ, and a random walk with drift model (RWD) with drift D. Student-t values are given in parentheses.

m	Men			Women		
	AR1		RWD	AR1		RWD
	D	ρ	D	D	ρ	D
2	0.0919 (0.1)	0.9876 (14.1)	−0.1354 (−11.3)	0.0028 (0.0)	0.9873 (11.3)	−0.0755 (−10.2)
3	0.2807 (0.1)	0.9877 (13.5)	−0.4671 (−12.4)	0.1748 (0.1)	0.9863 (12.1)	−0.2547 (−6.4)
4	−0.1118 (−0.1)	0.9859 (12.9)	0.2596 (6.8)	−0.0054 (−0.0)	0.9882 (13.4)	0.0468 (15.6)
5	−0.5283 (−0.2)	0.9873 (12.4)	0.4831 (10.9)	−0.8042 (−0.2)	0.9870 (11.3)	0.5880 (9.7)
6	−0.0795 (−0.7)	0.0745 (0.2)	0.0735 (0.5)	0.0429 (0.2)	0.9765 (9.4)	0.0367 (3.6)

never significantly different from zero, while the autoregressive parameter ρ has an estimate that is almost equal to one in all but one case (men, $m = 6$). This suggests that a RWD is a good choice for the time index, provided that one accepts that error terms of this RWD may be correlated. Estimation using OLS would lead to overly small standard errors, meaning that standard t-tests would not apply. Therefore, the results in Table 1 were estimated by maximum likelihood. For all nine cases, we note that the drift estimates have the signs that we would expect, based on Fig. 3. For the case of men, $m = 6$, a simple random walk is an appropriate choice.

3.2. Brass

We fitted the Brass-RWD model in Eq. (9) by means of OLS to empirical values of the logit-fractions ξ_2 to ξ_6, by age group (15–19, 20–24, …, 90–94, 95+) and sex, for the years

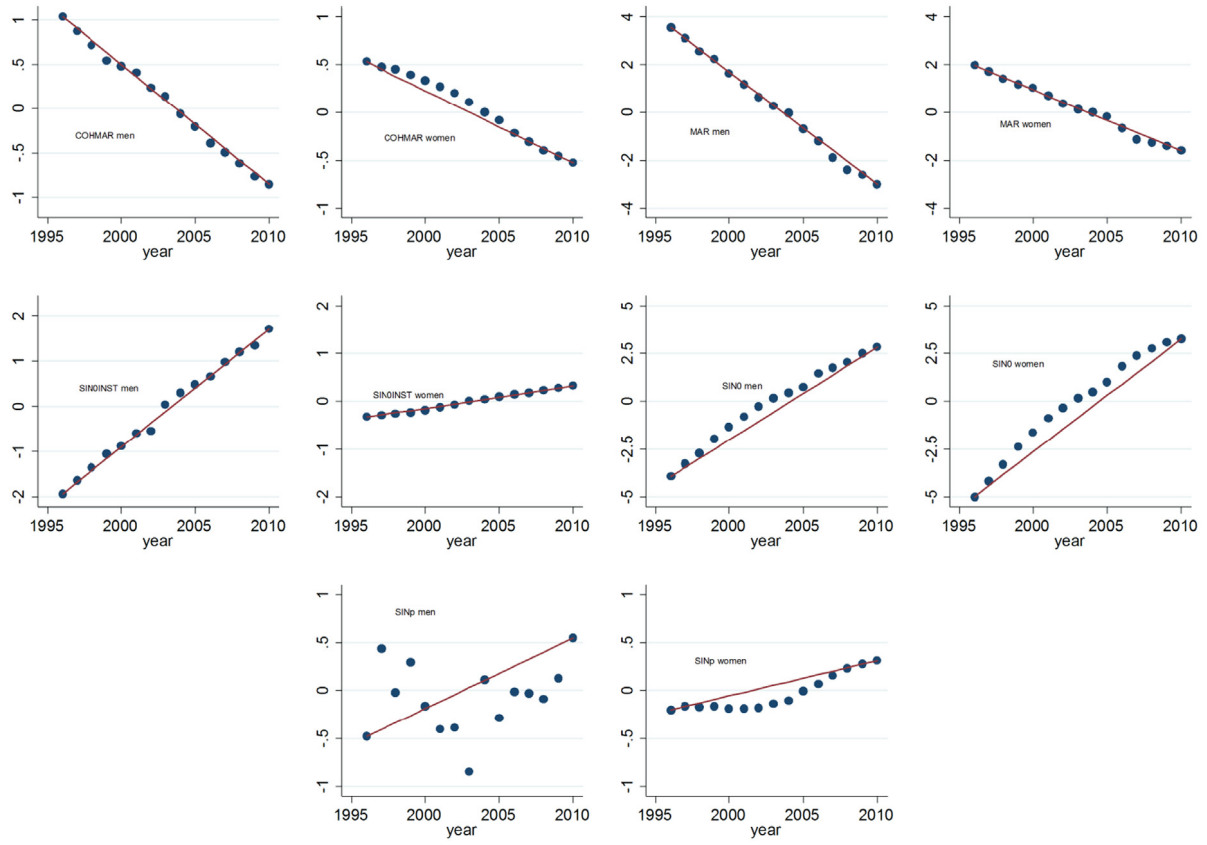

Fig. 3. Fit of the time index $k(t)$ estimates (dots) to a random walk with drift process (solid line) for $m = 2$ (COHMAR) to $m = 6$ (SINp), men and women.

Table 2
Parameter estimates and coefficients of determination for the Brass model in Eq. (9).

m	a_m		b_m		R^2
	Estimate	t-value	Estimate	t-value	(%)
2	−0.0060828	−2.8	−0.0049615	−2.3	3.0
3	−0.034871	−4.1	0.0091232	2.7	2.4
4	−0.0045308	−1.4	0.0097361	4.2	2.1
5	0.0483856	14.0	−0.0097381	−7.3	8.1
6	0.014209	3.6	0.0089384	1.0	1.6

1996–2010. As the parameter estimates for each type of fraction $\xi_m(x)$ differed little between men and women, and differences were not significant in most cases, we fitted the model for the two sexes combined. Table 2 gives the results. The estimate of b_6 is not significantly different from zero (at the 5% significance level). Thus, in this case the RWD contains a drift that is independent of age; cf. Eq. (9) above. Very little of the variation in $\Delta\xi$ is explained by this model; note the R^2-values of 2%–8%. Note however that this concerns first differences in ξ, which are quite small compared with the actual ξ-values. Indeed, one-year-ahead in-sample predictions of $\xi(x, t)$ – for instance $E[\xi(x, 1997)]$ given the value of $\xi(x, 1996)$ – agree with the data much better (not shown here). Out-of-sample predictions until 2020, 2030, and 2040 are presented in the next section.

4. Predictions

In Fig. 4, we show shares α_j for women in selected household positions, namely COH ($j = 3$), MAR ($j = 4$), SIN0 ($j = 2$), and COH + MAR combined ($\alpha_3 + \alpha_4$) for the years 1996, 2010, 2020, 2030, and 2040, based on in-sample and out-of-sample predictions of the LC-RWD model. While the age profile for cohabiting women seems reasonable for future years, that for women who live with a marriage partner becomes increasingly skewed by 2040. By then, women younger than age 40 are very unlikely to be married, while the opposite seems to be the case for those around age 70. In order to check for a possible substitution between household positions COH and MAR, one can also inspect the combined plot for these two positions (COH + MAR). This shows an unrealistic distortion in the age profile occurring at age 60, due to the estimates $\hat{b}(x)$ (not shown here) for this combined household position ($m = 2$; see Section 2.2), which fall from positive values for ages up to 55, to negative values for ages 60–90. This is a consequence of the cross-over in the age profile between 1996 and 2010 that was noted in connection with Fig. 2. The time index $k(t)$ starts with a negative value in 2010, and falls further for future years because its drift is negative (−0.0755; see women, $m = 2$ in Table 1). As a result, the product $b(x) \cdot k(t)$ pushes the age profile for these women *down* for ages up to 55, but pushes it *up* for ages 60–90 (the

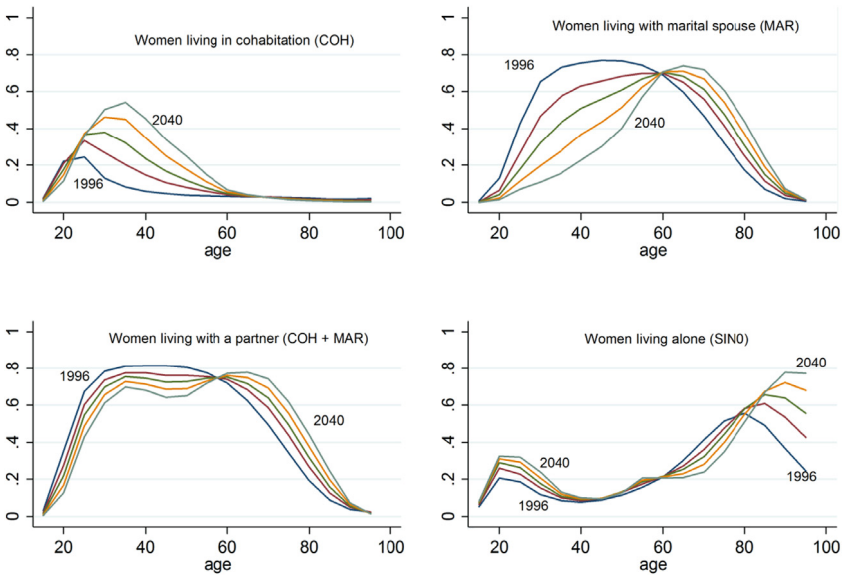

Fig. 4. Shares of women in selected household positions; observed values in 1996 and 2010, Lee–Carter random walk with drift predictions in 2020, 2030, and 2040.

estimate for $b(x)$ at age $x = 95$ is slightly positive again). A similar distortion can be seen in the age profile for married women. Again, the $b(x)$-estimates (not shown here) change sign: they are positive for ages 15–55, and negative for ages 60 and over. Because of the hierarchy, these distortions propagate to cases $m = 4$, 5, and 6. Indeed, in spite of the strictly positive $b(x)$-estimates (not shown here) for women who live alone (SIN0), some odd twists in the age profile are visible for the years 2030 and 2040. We do not show any results for men here, but their age patterns display distortions similar to those for women.

More generally, the LC-RWD model may lead to distorted age patterns when the estimates of $b(x)$ change signs. In that case, twists in the age patterns may occur: for positive values of $k(t)$, $b(x)$-estimates that are positive at some ages and negative at others indicate that the dependent variable tends to rise at some ages while falling at others. For negative $k(t)$-values, the situation is reversed. Lee and Miller (2001) argue that, for the case of mortality, $b(x)$ does not change sign in practice, as long as the model is fitted over a fairly long period. This is because, in most countries of the world, mortality has declined at all ages in the long run; however, this is not the casein our household application. Would other identification constraints for $b(x)$ help? One possibility could be to require $\Sigma_x b^2(x) = 1$ (Girosi & King, 2008; Wilmoth, 1993), or to estimate the model using the restriction that all $b(x)$ must be non-negative. This is not pursued here, because the true underlying cause is not the choice of restrictions, but the particular form of the LC-model. A singular value decomposition of the matrix $M(x, t) = \{\ln(m(x, t) - \hat{a}(x))\}$ results in three unique matrices, U, Σ and V^T. Thus, if one requires $b(x) \geq 0$ for all x, this will alter the estimates of $k(t)$, because the product of these two still has to be equal to the rank-1 approximation of $M(x, t)$. As a result, the $k(t)$ estimates will probably no longer resemble a straight line, which will make it more difficult to extrapolate them to future years.

We now turn to shares computed by means of the Brass-RWD model; see Fig. 5. Unlike the LC-RWD extrapolations in Fig. 4, the age profile of women who live with a marital spouse (MAR; $j = 4$) does not show any unrealistic twists. Except for the distortions in the age profile, the Brass extrapolations show roughly the same patterns as the Lee–Carter extrapolations in the previous figure.

Note that the Brass method predicts a continuous fall in the shares of MAR at ages 20 to 60, in line with historical observations. For ages 65 onwards, the shares for MAR increase between 1996 and 2010, due to falling mortality rates, and hence, the postponement of widowhood. However, unlike the LC-RWD model, the Brass-RWD model does not pick up this time trend in the shares for MAR, cf. Fig. 4. Extending the Brass model with a third parameter to represent the shape better at some ages (cf. Ewbank, Gomez de Leon, & Stoto, 1983, for the case of mortality) might solve the issue, but we have not done that here, because, with only 16 years of data, we wanted to keep the number of parameters to a minimum.

Fig. 5 reveals another issue. The *per annum* change in age-specific shares for positions COH and MAR appears to be stronger in the observation period 1996–2010 than over the period 2010–2040. The reason for this is that the shares of MAR for women younger than 40 years of age fell more steeply over the first few years of the observation period, roughly the period 1996–2000, than during the remaining years, roughly 2001–2010. The same can be said of the corresponding fractions for $m = 3$ in the logit scale. Thus, there are a few years with relatively large values of $\xi_3(x, t)$, and many years with smaller values. For the Brass-RWD model, the standard profile $\xi_3^S(x)$ for the fractions $\xi_3(x, t)$ is taken as the average value of $\xi_3(x, t)$ over the years 1996–2010. Therefore, this average is smaller than one would expect from a simple comparison of the profiles for the two years 1996 and 2010. These relatively small values of the standard, together with the estimates for model

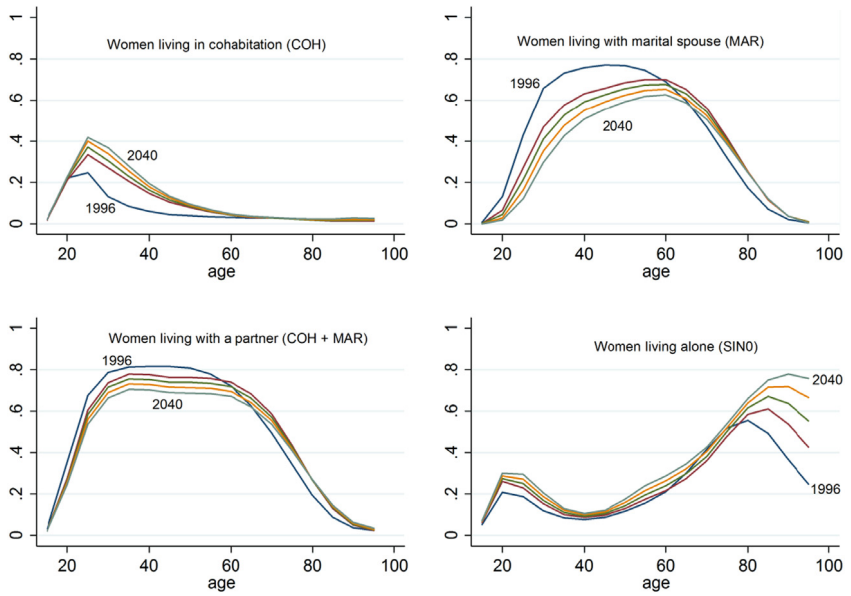

Fig. 5. Shares of women in selected household positions; observed values in 1996 and 2010, Brass random walk with drift predictions in 2020, 2030, and 2040.

parameters a and b, define the annual increments in the predictions of $\xi_3(x, t)$; see Eq. (10). Note that this feature of tempo changes in the development of age profiles is *not* observed in the corresponding predictions from the LC-RWD model; cf. Fig. 4. This is because the annual increments in LC-RWD predictions are determined by the product $D \cdot b(x)$ (see Eq. (8)), *not* by the standard profile in the LC-RWD model, namely $a(x)$.

5. Conclusions and discussion

We have formulated two different time series models for changes over time in the age profiles of six household positions that men and women can occupy at any point in time. These household positions are living alone, cohabiting, living with spouse, lone parent, living in some other private household position, and living in an institution. Preserving the characteristic features of such age profiles while accounting for their changes over time is an important task when predicting the household positions of individuals into the future. Both models are random walk with drift (RWD) models, in which the year-on-year step for the household parameter consists of a certain fixed term (the drift) plus a normally distributed error term with zero expectation. We tested two different versions of the RWD model, namely one based on the Lee–Carter model (LC-RWD model), which was originally developed for age patterns of mortality, and the other starting from the Brass relational model, which also stemmed from mortality analyses (Brass-RWD model). When the models are applied to data from the Netherlands for the period 1996–2010, we find that the Brass-RWD model predicts more realistic age profiles of household parameters than the LC-RWD model. This is because of the particular form of the LC-RWD model. When empirical age profiles show a cross-over (a downward trend for some ages, and an upward trend for others), the LC-RWD model is not appropriate, as one of its parameters, namely the $b(x)$-vector, may have both positive and negative values. This may lead to extrapolated age patterns for future years that are distorted strongly. This distortion becomes more severe for more distant years, because the $b(x)$ values are multiplied by falling or increasing values of a second vector, namely the time index $k(t)$ of the model. Extrapolations obtained using the Brass-RWD model did not show these kinds of distortions in our application.

The Brass-RWD model may be viewed as a two-parameter model (parameters a and b, see Eq. (9)), which makes it more flexible than the one-parameter LC-RWD model (parameter D in Eq. (5)). Thus, the fact that the Brass method works better should not come as a surprise. Unless the time series data being modeled and projected are predominantly linear, the two- parameter RWD will almost certainly outperform the one-parameter RWD.

A second, more general reason to prefer the Brass-RWD model to the LC-RWD model is the fact that the parameters of the former can be estimated very simply by ordinary least squares regressions. Closed form expressions for standard errors of the estimates, and standard deviations of the forecast errors, can be obtained using many statistical packages. On the other hand, the LC-RWD model is estimated by a singular value decomposition. While this decomposition has been included in statistical packages as well, it is not straightforward to obtain standard errors of the parameter estimates. Lee and Carter (1992), and many others since, used bootstrapping to simulate these errors. They also simulated prediction intervals, but they had to base their simulations on a number of assumptions. Hence, as they admit, the intervals that they obtained

may have been too narrow. An alternative in the context of mortality, which was suggested originally by Wilmoth (1993), is to assume that deaths counts follow a Poisson distribution, with the Poisson parameter modelled by means of the LC-model. Next, one can use the maximum likelihood method to estimate the LC parameters and the corresponding standard errors; see Chapter 5 of Alho and Spencer (2005) for a general treatment. One issue here is that, under this model, the exposure time $E(x, t)$ for the population at risk of dying in the age-time interval depends on the LC parameters, which leads to a complicated likelihood function. However, in one application, Alho and Nyblom (1997) found assuming an exposure time that is independent of the model parameters to have very little effect.

Clearly, in certain specific cases, the effect that we call "distortion" could be the result of cohort effects. For instance, the age profile of women who live with their spouses could display a maximum at a certain age in a given year because these women have a more positive view on marriage than women from other birth cohorts. If this cohort effect lasts, the top in the profile will move to higher ages for later years. While cohort effects of this type cannot be excluded in general, we do not believe that they cause the distorted age patterns for shares MAR and SIN0 in Fig. 4. If this were the cause, the top of the curve would shift to the right by ten years of age for every ten-year period. Implementing any explicit cohort effect in the models would require a standard profile for birth cohorts, in addition to one for periods. We have not done this here, because we wanted to keep the number of parameters to a minimum due to our relatively short time series.

As Section 4 showed, the Brass model has its limitations. In our application, it was not able to model the postponement of widowhood among married women. This problem could be solved by adding one or more parameters to the model, thus making it more flexible.

Another issue is that of coherence between men and women. In the observed data, there is a close correspondence between the numbers of men and women in household types COH and MAR. The numbers are not exactly equal, due to partnership formation and marriage across international borders, same-sex couples, and errors in the registration, but they are close. However, this coherence is lost when we predict shares for cohabiting and married men and women separately. When the predicted shares are combined with the results of a forecast of the population broken down by age and sex, this may lead to very different predictions for the numbers of men and women in household positions COH and MAR.

To sum up, this paper shows how data reduction techniques stemming from mortality analysis can be used to describe and project household dynamics. More specifically, we show that the Brass method has considerable advantages over LC for this particular data set: the resulting age patterns for men and women in various household positions look more realistic, and the model parameters are easier to estimate. At the same time, we have to acknowledge its limitations: some time trends are not captured, the model does not include cohort effects, and the coherence between men and women who live in a partnership is not taken into account.

Acknowledgments

Useful comments from Coen van Duin and two anonymous reviewers are gratefully acknowledged.

References

Alho, J. M., & Nyblom, J. (1997). Mixed estimation of old-age mortality. *Mathematical Population Studies*, *6*, 319–330.
Alho, J. M., & Spencer, B. (2005). *Statistical demography and forecasting*. New York: Springer.
Booth, H. (1984). Transforming the Gompertz for fertility analysis: the development of a standard for the relational Gompertz. *Population Studies*, *38*(3), 495–506.
Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, *22*(3), 547–581.
Box, G., & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden Day.
Brass, W. (1971). On the scale of mortality. In W. Brass (Ed.), *Biological aspects of demography* (pp. 69–101). London: Taylor and Francis Ltd..
Burch, Th. (1995). Theories of household formation: progress and challenges. In E. Van Imhoff, A. Kuijsten, P. Hooimeijer, & L. Van Wissen (Eds.), *Household demography and household modeling* (pp. 85–108). New York: Plenum Press.
Christiansen, S., & Keilman, N. (2013). Probabilistic household forecasts based on register data: the case of Denmark and Finland. *Demographic Research*, *28*, 1263. 06/2013.
Coale, A., & Trussell, J. (1974). Model fertility schedules: variations in the age structure of childbearing in human populations. *Population Index*, *40*(2), 185–258.
De Beer, J. (2011). A new relational method for smoothing and projecting age specific fertility rates: TOPALS. *Demographic Research*, *24*, 409–454.
De Beer, J. (2012). Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research*, *27*, 543–592.
Ewbank, D., Gomez de Leon, J., & Stoto, M. (1983). A reducible four-parameter system of model life tables. *Population Studies*, *37*(1), 105–127.
Girosi, F., & King, G. (2008). *Demographic forecasting*. Princeton: Princeton University Press.
Holmans, A. (2012). *Household projections in England: Their history and uses*. England: Paper, Cambridge Centre for Housing & Planning Research, University of Cambridge.
Lee, R. D., & Carter, L. (1992). Modeling and forecasting the time series of US mortality. *Journal of the American Statistical Association*, *87*, 659–671.
Lee, R., & Miller, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography*, *38*(4), 537–549.
Lesthaeghe, R. (1995). The second demographic transition in Western countries: an interpretation. In K. O. Mason, & A.-M. Jensen (Eds.), *Gender and family change in industrialized countries* (pp. 17–62). Oxford: Clarendon Press.
Preston, S., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modelling population processes*. Oxford: Blackwell Publishers.
Van de Kaa, D. (1987). Europe's second demographic transition. *Population Bulletin*, *42*, 1–59.
Van Imhoff, E., & Keilman, N. (1991). *LIPRO 2.0: An application of a dynamic demographic projection model to household structure in the Netherlands*. Amsterdam, Berwyn, PA: Swets and Zeitlinger Publishers.
Verdon, M. (1998). *Rethinking households*. London: Routledge.
Wilmoth, J. R. (1993). Computational methods for fitting and extrapolating the Lee–Carter model of mortality change. In *Technical Report Department of Demography*. University of California at Berkeley.
Zeng, Y., Wang, Z., Ma, Z., & Chen, C. (2000). A simple method for estimating α and β: An extension of Brass Relational Gompertz Fertility model. *Population Research and Policy Review*, *19*(6), 525–549.

Nico Keilman is Professor of Demography at the Department of Economics, University of Oslo.

He has more than 25 years of experience in research on population and households. His academic interests include population forecasting, modelling marriage and household dynamics, and mathematical demography.