Stochastic household forecasts by coherent random shares predictions

Nico Keilman

Coen van Duin

Version 4 November 2012

Abstract

We compute a stochastic household forecast for the Netherlands by the random share method. Time series of shares of persons in nine household positions, broken down by sex and five-year age group for the years 1996-2010 are modelled by means of the Hyndman-Booth-Yasmeen product-ratio variant of the Lee-Carter model. This approach reduces the dimension of the data set by collapsing the age dimension into one scalar. As a result, the forecast task implies predicting two time series of time indices for each household position for men and women. We model these time indices as a Random Walk with Drift (RWD), and compute prediction intervals for them. Prediction intervals for random shares are simulated based on the Lee-Carter model. The random shares are combined with population numbers from an independently computed stochastic population forecast of the Netherlands.

Our general conclusion is that the method proposed in this paper is useful for generating errors around expected values of shares that are computed independently. In case one wishes to use this method for computing the expected values for the household shares as well, one has to include cohort effects in the Lee Carter model. This requires long time series of data.

1. Introduction

Since 1999, Statistics Netherlands has published bi-annual updates of a stochastic household forecast; see Alders (1999, 2001) and De Beer and Alders (1999). Alders and De Beer combined a stochastic population forecast with random shares. The shares distribute the population probabilistically over six household positions: individuals could live as a child with parents, live alone, live with a partner, as a lone parent or in an institution, or belong to another category. For instance, the authors computed the (random) number of lone mothers aged 40 in 2015 as the product of two other random variables, namely the number of women aged 40 in 2015 and the share of 40-year old women who live as a lone mother in 2015. Expected values for population variables and for the shares for specific household positions were obtained from observed time series, but the statistical distributions that were assumed for the shares were based on intuitive reasoning.Perfect correlations across age and sex were assumed for the mortality rates, fertility rates and migration numbers in the stochastic population forecasts, as well as for the random shares. In addition the authors assumed perfect correlation in the time dimension for the random shares.

Our aim is to improve on this by computing a probabilistic household forecast for The Netherlands that is based on empirically specified uncertainty parameters, instead of intuitive values. We will use data for the period 1996-2010 to estimate time series models for the shares that distribute the population over household positions. The estimated time series models allow us to compute predictive distributions for these shares, and to evaluate correlations in the shares across age and sex.

2. Method

We want to assess the uncertainty in predicted household shares. Write V(j,x,s,t) for the number of people in household position j=1,2, . . . who are in age x=0,1, . . . and have sex s=1 or 2, at time t=0,1,2, ... Aggregating over position, we obtain the population W(x,s,t) = Σj V(j,x,s,t) of age x and sex s at time t. The share of household position j is α(j,x,s,t) = V(j,x,s,t)/W(x,s,t) = α_j(x,s,t).

2.1 Predictions of the shares

We have data for men (s=1) and women (s=2) in the Netherlands broken down by age group (x=1 for age 0-4, x=2 for age 5-9, …, x= 20 for age 95+) on their household positions as of 1 January of the years 1996 (t=1), 1996 (t=2), …, 2010 (t=15). We distinguish the following nine household positions (j=1,2,…,9):

CHLD j=1: dependent child living with parents

SIN0 j=2: living in one-person household

COH0 j=3: living in unmarried cohabitation, no children

COH+ j=4: living in unmarried cohabitation, one or more children

MAR0 j=5: living with marital spouse, no children

MAR+ j=6: living with marital spouse and one or more children

SIN+ j=7: lone parent

OTHR j=8: other position in private household, for instance member of multiple family household, person living with non-family related individuals, homeless

INST j=9: living in an institution

No age restrictions have been imposed on persons who have a certain household position. In particular, children (CHLD) and lone parents (SIN+) can be of any age. In practice, predicted numbers of persons aged 85, say, with positions CHLD or SIN+, will not be interpreted as such, but should be assigned to a different position, for instance to the group of other. Moreover, we have ignored persons aged younger than 15 in the following positions: SIN0, COH0, COH+, MAR0, MAR+, and SIN+.

For modelling 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 eight types of fraction to be modelled (all specific for age, sex and time). By construction, the eight fractions as listed below can be interpreted as representing stochastically independent conditional probabilities. Independence is an advantage when we predict the values of these random shares into the future. We have chosen two alternative specifications of the hierarchy. The first one is similar to the one used in previous work for Denmark, Finland, and Norway (Alho and Keilman 2010; Christiansen and Keilman 2012). The following shares were used:

<u>Specification 1</u>

1. The total share of SIN0, MAR0, and MAR+;

2. The relative share of MAR0 and MAR+ out of the total share of MAR0, MAR+, and SIN0;

3. The relative share of MAR0 out of the share of MAR0 and MAR+;

4. The relative share of COH0 and COH+ out of the total share of COH0, COH+, CHLD, SIN+, OTHR, and INST;

5. The relative share of COH0 out of the share of COH0 and COH+;

6. The relative share of CHLD out of the total share of CHLD, OTHR, SIN+, and INST;

7. The relative share of SIN+ out of the total share of SIN+, OTHR, and INST;

8. The total share of INST out of the total share of INST and OTHR.

The particular sequence 1-8 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. 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.

An alternative specification gives more weight to households with resident children than to those without. This leads to the following set of shares:

Specification 2

a. The total share of COH+, MAR+, and SIN+;

b. The relative share of MAR+ out of the total share of COH+, MAR+, and SIN+;

c. The relative share of COH+ out of the share of COH+ and SIN+;

d. The relative share of CHLD out of the total share of CHLD, SIN0, COH0, MAR0, OTHR, and INST;

e. The relative share of SIN0 out of the total share of SIN0, COH0, MAR0, OTHR, and INST;

f. The relative share of CoH0 out of the total share of COH0, MAR0, OTHR, and INST;

g. The relative share of MAR0 out of the total share of MAR0, OTHR, and INST;

h. The total share of INST out of the total share of INST and OTHR.

In what follows, we will refer to Specification nr. 1, unless stated otherwise.

Temporarily suppressing indices for age, sex, and time, the logit transform of the share of type 1 for Specification nr. 1 above is

(1) $\xi_1 = \text{logit}(\alpha_2+\alpha_5+\alpha_6) = \log((\alpha_2+\alpha_5+\alpha_6)/(1-\alpha_2-\alpha_5-\alpha_6))$.

For the logit transformed shares of types (2)-(8) we find

(2) $\xi_2 = \text{logit}((\alpha_5+\alpha_6)/(\alpha_2+\alpha_5+\alpha_6)) = \log((\alpha_5+\alpha_6)/\alpha_2)$

(3) $\xi_3 = \text{logit}(\alpha_5/(\alpha_5+\alpha_6)) = \log(\alpha_5/\alpha_6)$

(4) $\xi_4 = \text{logit}((\alpha_3+\alpha_4)/(\alpha_1+\alpha_3+\alpha_4+\alpha_7+\alpha_8+\alpha_9)) = \log((\alpha_3+\alpha_4)/(\alpha_1+\alpha_7+\alpha_8+\alpha_9))$

(5) $\xi_5 = \text{logit}((\alpha_3/(\alpha_3+\alpha_4)) = \log(\alpha_3/\alpha_4)$

(6) $\xi_6 = \text{logit}(\alpha_1/(\alpha_1+\alpha_7+\alpha_8+\alpha_9)) = \log(\alpha_1/(\alpha_7+\alpha_8+\alpha_9))$

(7) $\xi_7 = \text{logit}(\alpha_7/(\alpha_7+\alpha_8+\alpha_9)) = \log(\alpha_7/(\alpha_8+\alpha_9))$

(8) $\xi_8 = \text{logit}(\alpha_9/(\alpha_8+\alpha_9)) = \log(\alpha_9/\alpha_8)$

Similar logit transformed shares can be derived for Specification nr. 2.

This way, eight stochastically independent time series (given age and sex) were constructed. With two sexes and 20 age groups, the theoretical number of time series is 320. In practice, we have 284 series, because children younger than age 15 can be in household positions CHLD and INST only (in addition to OTHR, which is the reference category).

In the logit scale, the time series show approximately a linear time trend (details are available upon request). We wish to extrapolate these time series for each household position, thereby preserving the age pattern and the coherence between men and women. In case we would extrapolate each time series separately, unreasonable age patterns may arise for predicted shares, in particular in the long run. For example, the age pattern of women of MAR0 (ξ_3, in other words the log odds of being married with no children in the household, compared to being married with one or more children) has changed over time. For young and middle-aged women (aged 25-49) the odds have fallen, while they have increased for elderly women (aged 65+). Fewer children in the household as a result of the decrease in fertility after the baby boom explain the time trend in ξ_3. If one would extrapolate each age group of ξ_3 separately, an unrealistic age pattern may arise, with unreasonably low values for young and middle-aged women may arise, and very high values for elderly women.[1]

We have used the Hyndman-Booth-Yasmeen product-ratio variant of the Lee-Carter model (LC model) to preserve the age patterns and the coherence between men and women. Originally developed for predicting age-specific mortality rates, the LC model assumes that the logarithm of the rate m(x,t) for age x during year t can be written as

$$\ln(m(x,t)) = a(x) + b(x).k(t) + \varepsilon(x,t).$$

The rate m(x,t) in log-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). ε(x,t) is an error term with the usual properties. Lee and Carter modelled the estimated time series k(t)

[1] Christiansen and Keilman (2012) experimented with separate time series for each age in a similar study based on data for Denmark and Finland. Their results for Finland in 2037 suggested that only around 60% of the population in the age group 15-19 would live with their parent(s), and hardly any in the age group 20-24. In Denmark it all but extinguished the share of elderly living in other private households.

as a RWD, and used the extrapolated k(t) values to predict age-specific mortality rates for future years.

Hyndman et al. (2012) noted that the LC model, when applied to mortality rates of men and women separately, may lead to unreasonable differences between male and female mortality in the long run, for instance a cross-over, or an increasing gap in terms of the life expectancy of the two sexes. Therefore they proposed to apply an LC model to each of the following transformed rates:

$$p(x,t) = \sqrt{[m(x,1,t).m(x,2,t)]} \text{ and } r(x,t) = \sqrt{[m(x,1,t)/m(x,2,t)]},$$

where m(x,1,t) and m(x,2,t) are the death rates for men and women, respectively. The product and ratio transformations defined above preserve the coherence between mortality of men and women, because the product and ratio will behave roughly independently of each other. Hyndman et al. (2012) argue that on the log-scale, these are sums and differences which are approximately uncorrelated.

Given predicted values of p(x,t) and r(x,t), predicted mortality rates for men and women are found as

$$m(x,1,t) = p(x,t).r(x,t) \text{ and } m(x,2,t) = p(x,t)/r(x,t).$$

The shares ξ_1 to ξ_8 in the logit scale defined above can be interpreted as log-odds. For instance, $\xi_2 = \text{logit}((\alpha_5+\alpha_6)/(\alpha_2+\alpha_5+\alpha_6))$ can be interpreted as the log of the conditional odds $(\alpha_5+\alpha_6)/\alpha_2$. Write this odds-value as $\beta_2=\exp[\xi_2]$, and similarly for ξ_1 and ξ_3 to ξ_8. This defines odds-values $\beta_k = \beta_k(x,s,t)$ (k=1,2,...8). Next, following Hyndman et al., we defined the following products and ratios:

$$p_k(x,t) = \sqrt{\{\beta_k(x,1,t).\ \beta_k(x,2,t)\}} \text{ and}$$

$$r_k(x,t) = \sqrt{\{\beta_k(x,1,t)/\beta_k(x,2,t)\}}, \text{ or equivalently}$$

$$(9) \quad p_k(x,t) = \exp[(\xi_k(x,1,t)+\xi_k(x,2,t))/2] \text{ and}$$

$$(10) \quad r_k(x,t) = \exp[(\xi_k(x,1,t)-(\xi_k(x,2,t))/2)],$$

where $\xi_k(x,s,t)$ denotes the share in the logit scale for household position k=1,2,...8, sex s (s=1 for men and s=2 for women), age x and year t. Next we estimated LC models for the products $p_k(x,t)$ and the ratios $r_k(x,t)$ as

(11) $\ln(p_k(x,t)) = a_{pk}(x) + b_{pk}(x).k_{pk}(t) + \varepsilon_{pk}(x,t)$, and

(12) $\ln(r_k(x,t)) = a_{rk}(x) + b_{rk}(x).k_{rk}(t) + \varepsilon_{pk}(x,t)$,

predicted $k_{pk}(t)$ and $k_{rk}(t)$ into the future to find future values of $p_k(x,t)$ and $r_k(x,t)$, and found $\xi_k(x,1,t)$ and $\xi_k(x,2,t)$ as $\ln[p_k(x,t).r_k(x,t)]$ and $\ln[p_k(x,t)/r_k(x,t)]$, respectively. Finally, we transformed predicted $\xi_k(x,s,t)$ back to shares $\alpha_k(x, s, t)$ as follows (suppressing x, s, and t):

(13) SIN0: $\alpha_2 = \exp[\xi_1]/\{(1+\exp[\xi_1])((1+\exp[\xi_2])\}$

(14) MAR+: $\alpha_6 = \alpha_2.\exp[\xi_2]/(1+\exp[\xi_3])$

(15) MAR0: $\alpha_5 = \alpha_2.\exp[\xi_2]-\alpha_6$

(16) COH+: $\alpha_4 = (1-\alpha_2-\alpha_5-\alpha_6).\exp[\xi_4]/\{(1+\exp[\xi_4])(1+\exp[\xi_5])\}$

(17) COH0: $\alpha_3 = \alpha_4.\exp[\xi_5]$

(18) CHLD: $\alpha_1 = (1-\alpha_2-\alpha_3-\alpha_4-\alpha_5-\alpha_6).\exp[\xi_6]/(1+\exp[\xi_6])$

(19) SIN+: $\alpha_7 = (1-\alpha_1-\alpha_2-\alpha_3-\alpha_4-\alpha_5-\alpha_6).\exp[\xi_7]/(1+\exp[\xi_7])$

(20) INST: $\alpha_9 = (1-\alpha_1-\alpha_2-\alpha_3-\alpha_4-\alpha_5-\alpha_6-\alpha_7).\exp[\xi_8]/(1+\exp[\xi_8])$

(21) OTHR: $\alpha_8 = 1-\alpha_1-\alpha_2-\alpha_3-\alpha_4-\alpha_5-\alpha_6-\alpha_7-\alpha_9$

Expressions (13)-(21) apply to Specification nr. 1. Expressions for the back-transformation from $\xi_k(x,s,t)$ back to shares $\alpha_k(x, s, t)$ for Specification nr. 2 are straightforward.

2.2 Prediction intervals for the shares $\alpha_k(x,s,t)$

The procedure outlined in Section 2.1 gives point forecasts for the shares, that is, predictions $E[\alpha_j(x,s,T+h)]$ (j=1,2,9) for h years into the future, based on data for the years t=1,2,...T. The prediction interval of the future share $\alpha_j(x,s,T+h)$ depends on its variance $Var[\alpha_j(x,s,T+h)]$. The variance of $\xi_k(x,s,T+h)$ (k=1,2,...8) is straightforward to compute. But because of the exponential transformation from ξ_k back to α_j in expressions (13) to (21), only approximate expressions for $Var[\alpha_j(x,s,T+h)]$ are known. Therefore we have used simulation to determine

the prediction intervals of the shares α_j. Before discussing this simulation procedure, we derive the variance of $\xi_k(x,s,T+h)$ first. We suppress variables for age and sex.

We have assumed a RWD model for the time indices $k_{pk}(t)$ and $k_{qk}(t)$ of the LC models (11) and (12). In particular, for the time indices k_{pk} of the products p_k we assumed

$$k_{pk}(t+1) = k_{pk}(t) + D_{pk} + \theta_{pk}(x,t+1),\ t=1,2,\ldots T.$$

Having estimated the drift D_{pk} and the variance $\sigma^2_{\theta pk}$ of the residuals, we can extrapolate the time index h years and write

$$k_{pk}(T+h) = k_{pk}(T) + h.D_{pk} + \theta_{pk}(x,T+1) + \theta_{pk}(x,T+2) \ldots + \theta_{pk}(x,T+h).$$

Given the usual assumptions about independence and homoscedasticity, and taking $k_{pk}(T)$ as known, the variance of $k_{pk}(T+h)$ is

$$Var[k_{pk}(T+h)] = h^2.\sigma^2_{\delta pk} + h.\sigma^2_{\theta pk},$$

where $\sigma^2_{\delta pk}$ is the variance due to estimation error in estimating the drift, and $\sigma^2_{\theta pk}$ is the innovation variance of the RWD.

The LC model in expression (11) is used to predict the product as follows:

$$\ln(p_k(x,T+h)) = a_{pk}(x) + b_{pk}(x).k_{pk}(T+h) + \varepsilon_{pk}(x,T+h),$$

which implies that the variance of the product equals

$$Var[\ln(p_k(x,T+h))] = b^2_{pk}(x)\{h^2.\sigma^2_{\delta pk} + h.\sigma^2_{\theta pk}\} + \sigma^2_{\varepsilon pk},$$

where $\sigma^2_{\varepsilon pk}$ is the variance of the error term of the LC model[2].

In obvious notation we find for the variance of the ratio

$$\mathrm{Var}[\ln(r_k(x,T+h))] = b^2_{rk}(x)\{h^2.\sigma^2_{\delta rk} + h.\sigma^2_{\theta rk}\} + \sigma^2_{\varepsilon rk}.$$

Since $\xi_k(x,1,t)=\ln[p_k(x,t).r_k(x,t)]$ and $\xi_k(x,2,t)=\ln[p_k(x,t)/r_k(x,t)]$, we find for the variances of the fractions ξ_k of men the following expression:

(22) $$\mathrm{Var}[\xi_k(x,1,T+h)] = \mathrm{Var}[\ln\{p_k(x,t)\} + \mathrm{Var}[\ln\{r_k(x,t)\}] =$$

$$= b^2_{pk}(x)\{h^2.\sigma^2_{\delta pk} + h.\sigma^2_{\theta pk}\} + \sigma^2_{\varepsilon pk} + b^2_{rk}(x)\{h^2.\sigma^2_{\delta rk} + h.\sigma^2_{\theta rk}\} + \sigma^2_{\varepsilon rk},$$

where we have assumed that for a given combination of x, k, and h the product and ratio are uncorrelated; see Hyndman et al. (2012).

The right-hand side can be decomposed into contributions from the residuals of the LC model, the RWD residuals, and the estimation error for the drift term of the RWD-process, both for the product and the ratio term, as follows:

<---------- product term --------> <---------- ratio term ---------->

$b^2_{pk}(x)\{h^2.\sigma^2_{\delta pk} + h.\sigma^2_{\theta pk}\} + \sigma^2_{\varepsilon pk} + b^2_{rk}(x)\{h^2.\sigma^2_{\delta rk} + h.\sigma^2_{\theta rk}\} + \sigma^2_{\varepsilon rk},$

<drift> <RWD res> <LC res> <drift> <RWD res> <LC res>

Expression (22) gives also the variance $\mathrm{Var}[\xi_k(x,2,T+h)]$ for women, because $\mathrm{Var}[\ln\{p_k(x,t)/r_k(x,t)\}] = \mathrm{Var}[\ln\{p_k(x,t)\} - \ln\{r_k(x,t)\}] = \mathrm{Var}[\ln\{p_k(x,t)\}] + \mathrm{Var}[\ln\{r_k(x,t)\}]$.

The simulation procedure is as follows.

1. Given k, x, s, and h, draw an error e from a Student t-distribution with T-2 degrees of freedom.

[2] This assumes that both $a_{pk}(x)$ and $b_{pk}(x)$ are non-random variables. In reality, when estimating $\mathrm{Var}[\ln(r_k(x,T+h))]$ by SVD, one estimates $a_{pk}(x)$ for a given x as the average value over time of the logarithmic value of the empirical rates. To treat $a_{pk}(x)$ as non-random is not unreasonable. For $b_{pk}(x)$ one could argue that its estimator is random. Expressions for its standard error are not known, but probably one could find the standard error by bootstrapping. We have not done that. Therefore the variances estimated here may be a bit too low.

2. Scale this error up by a factor √{Var[$\xi_k(x,1,T+h)$]}; see expression (22).

3. Add the point forecast E[$\xi_k(x,s,T+h)$].

4. Transform the result to the share $\alpha_j(x,s,T+h)$; see expressions (13) to (21).

5. Multiply this share with the size of the population sub-group with the corresponding combination of sex, age, and time in one realization of an independent stochastic population forecast.[3]

Repeat this procedure as many times as required.

In practice we did not draw one error e in Step 1, but a set of errors that are correlated across age groups (given k, s, and h). Following earlier work (Alho and Keilman 2010, Christiansen and Keilman 2012) we assumed an AR1 process for the errors in the age dimension, independently of k, s, and h. The correlation was estimated based on the residuals of the LC model.

3. Results

We have annual data on shares for 1 January of the years 1996 to 2010. The whole data set comprises nine household positions (see Section 2.1), men and women, and ages 0-4, 5-9, …, 90-94, and 95+. For both specifications of the hierarchy of household positions, we have estimated the LC model in expressions (11) and (12), and predicted the fractions $\xi_k(x,s,T+h)$ for men and women in five-year age groups for lead times h=5, 10, 15, ..., 30 and eight household positions k=1-8, based on a Random Walk with Drift model for the time indices $k_{pk}(t)$ and $k_{qk}(t)$. Since the jump-off year T is 2010, this corresponds to future years 2015, 2020, 2025, ..., 2040. We used stochastic simulation to predict 1000 fractions ξ specific for household position, age group, sex, and lead time. These fractions were transformed to shares α and multiplied with stochastic population numbers, again specific for age group, sex, and lead time. The latter population numbers stem from the official stochastic population forecast of Statistics Netherlands. See Carolina and Van Duin (2010) for details.

When we present and discuss our findings below, we will compare our point predictions for the shares with similar predictions for the year 2040 derived from the 2011-based official (deterministic) household forecast of Statistics Netherlands; see Van Duin and Stoeldraijer (2011). This forecast distinguishes the same nine household positions as we do, but it uses a methodology that differs from ours in two important respects (Van Duin and Harmsen, 2009). First, Statistics Netherlands uses a multistate cohort component model to predict the population broken down by the same nine household positions as in our case, but in addition by four marital statuses: never married, currently married, divorced, and widow(er). Thus household positions are specific not only for age and sex, but also for marital status. The

[3] This multiplication assumes independence between household shares and population numbers. Reasons why we think that this is a plausible assumption are given by Alho and Keilman (2010).

multistate approach is dynamic in the sense that it models events in terms of changes in household position and changes in marital status. Our approach is static: each future year it breaks the population down into nine household categories. The advantage of the dynamic (event-based) approach is that it takes account of the cohort progression of events. Our approach preserves changes in age specific shares period-wise only, not cohort-wise. A further advantage of the dynamic approach is that it allows the modeller to take links between certain groups of events into account. For instance, the number of men who enter the position MAR0 during a certain time interval must be equal to the number of women who do so. Similar consistency relationships can be formulated for the formation of cohabiting unions and for union (marital or consensual) dissolution, including one partner entering an institution while the other partner becomes a one-person household. In contrast, our extrapolated shares ignore constraints of this kind, because the household events remain a black box. A second difference is that parameter extrapolations are based on visual inspection, not on an explicitly formulated time series model.

3.1 General findings

Our findings can be summarized as follows.

The point predictions for the shares are more or less regular extrapolations of observed trends for the period 1996-2010. By and large, the resulting age patterns for the nine household positions of men and women are reasonable, but we note two important points. First, period effects are exaggerated in a few cases, while at the same time our procedure is unable to account for cohort effects. Second, there is no guarantee that we will obtain consistent numbers for future years of men and women who live as a couple. A third point is that there was little difference between the results of the two specifications of the hierarchy of household positions. Fourth, the uncertainty estimates that we obtain for future households in the Netherlands look reasonable.

Each of the four points will be illustrated below. The results that we report here apply to the first specification, unless stated otherwise.

Exaggerated period effects

In an initial set of predictions of the shares, we noticed that for a number of household positions, our approach would lead to much stronger shifts in the age patterns of the household shares than the approach used by Statistics Netherlands. This was the case for both specifications of the hierarchy of household positions. For instance, in prime reproductive ages, the chances of living with a marital spouse and one or more children (MAR+) would fall by about 20 percentage points over the next 30 years; see Figure 1 for women. The dominant place of this household position would be taken over by COH+, where we saw an increase by about 40 percentage points for young adults (Figure 2). These changes were about three times as strong as those foreseen by Statistics Netherlands. The difference

with our results is explained by Statistics Netherlands' assumption that the fall in marriage rates observed in the past has come to an end: the total period probability of first marriage, which has declined since 1970, is assumed to remain near the current level (70% for women, 65% for men) as the downward trend seems to have stopped around 2002 and prospective surveys show an increase in the proportion among the young who expect to marry (Van Duin and Stoeldraijer 2011 p. 60). Our time series approach extrapolates the effects of this decline on the share of married couples into the future. Thus more cohabiting couples would remain unmarried, compared to the results of Statistics Netherlands. For the household positions that include married and cohabiting persons we have attenuated, in both specifications, the time trends of the respective Random Walks with Drift. More specifically, for k equal 1, 2, 3, 4, and 5, we extrapolated the RWD of the product time indices $k_{pk}(t)$ not thirty, but only ten years into the future, after which each of these time indices were kept constant during the remaining twenty years. This procedure does not apply to ratio time indices $k_{rk}(t)$, which describe the relationship between men and women.

Figures 3-9 give point predictions for selected shares, including the fix for married and for cohabiting persons. Comparing Figures 5 and 7 with Figures 1 and 2, respectively, one notes that the large shifts over time in the age pattern of shares for COH+ and MAR+ are much less than in the initial extrapolations. Future shares for household position MAR0 are still a bit high, particularly those for women. Figure 3 for persons who live alone demonstrates clearly that the time series approach is not able to model cohort progression: a strong increase in the chances of living alone of men and women in their forties during the years 1995-2005 should show up as a similar increase for persons in their fifties during 2005-2015. Indeed, the cohort component approach of Statistics Netherlands does account for such cohort progression. Note that Statistics Netherlands predicts a bulge in the chances for persons in their sixties to live as COH0. This reflects the empty nest phase for these persons, but it is not yet visible in the historical data. The consistency algorithm of SN's model includes an explicit link between young adults who leave the parental household, and middle-aged persons who become SIN0, COH0, or MAR0 when the last child leaves them.

For the elderly who live in an institution (Figure 9) we note a similar decrease as Statistics Netherlands predicts. The downward trend is explained by the tendency of elderly to remain living on their own, rather than in an institution. This trend is visible in the historical data, and it is likely to continue because the capacity of institutions for the elderly cannot keep up with the pressure of an ageing population.

When we predict falling shares of elderly who are institutionalized, this implies that chances are increasing that these persons will live alone or with a partner. This is indeed the case; see position SIN0 for women (Figure 3) and, much less so, position MAR0 for men (Figure 6).

Middle aged men see a small increase in their chances of living alone, as a continuation of the trend in union dissolution (Figure 3). This trend is clearly visible in the time series of the relevant share since 1995. Statistics Netherlands assumes a stabilization of divorce and separation propensities, and hence predicts lower chances for men aged 40-60 to live alone in 2040 than we do. Cohort progression is clearly visible in the pattern predicted by Statistics

Netherlands for men and women aged around 65 who live alone. The LC model does not include a cohort effect and hence such cohort progression does not show up in our results. This problem could be solved by adding to the LC model in expression (11) an extra component that takes cohort effects into account. However, since the time series is short (1996-2010), we wish to keep the number of parameters to a minimum, and therefore we have not added such an extra component.

For women aged 60-85 the chances of living alone will fall slightly as a consequence of improved survival of married men; compare also Figure 6 for women.

Qualitatively speaking, our predicted age patterns look reasonable. One important exception was the pattern for lone mothers in Figure 8. Our initial extrapolations showed an odd twist emerging over the years in the age schedule for lone mothers aged 25-40. This is explained by temporarily high fertility levels of young girls with Antillean or Surinamese background. This phenomenon has become much weaker in recent years. To reduce its effects we have modified the values of the b(x) parameters for the relevant groups in such a way that the age pattern in Figure 8 looks reasonable, although a slight effect is still visible. The increase over time in chances for lone mothers aged 45-55 are explained by increasing union dissolution risks. It coincides with an increasing chance for men in similar age groups to live alone, as noted earlier. For lone fathers, the changes over time are minimal.

Inconsistent numbers for men and women who live as a couple

Predicted numbers of men and women who live as a couple (COH0, COH+, MAR0, or MAR+) were not consistent. Small differences are acceptable, because the observed numbers of men and women in these household positions are not entirely consistent either: some persons live with a partner of the same sex, or with a partner who is not (yet) entered into the population register. The latter may be caused by registration backlog after a move to a new address, by specific registration rules for immigrants and tourists, by administrative errors and the like. However, the inconsistencies that we encountered for future years are too large to be plausible. We computed, for each of the 1000 simulations, total numbers of men and women who live as a couple, as well as the sex ratios (number of men per 100 women). Table 1 below shows average values of the sex ratios across 1000 simulations in 2040.

Table 1. Average sex ratios for the number of persons in four household positions, 2040

COH0	COH+	MAR0	MAR+
90.1	85.8	122.8	93.1

One may argue that the inconsistency between men and women who live as a couple is caused by the drift in the RWD model for the ratios r_k in expression (19). Therefore we removed the drift terms of the Random Walks for fractions ξ_1, ξ_3, ξ_4, and ξ_5. The results were disappointing. For instance, for the year 2040 we found little or no change in the sex ratios for COH0 (86.5) and MAR0 (119.6), while those for COH+ (130.4) and MAR+ (115.3) showed strong imbalances in the opposite direction. We obtained similar imbalances when in addition, standard deviations of the innovation terms of the RWD process were set to zero (in other words, when we assumed $\sigma_{\theta rk} = 0$ for k = 1,3,4, and 5 so that the ratios were kept constant 30 years into the future).

Similar results for the two specifications of the hierarchy of household positions

The second specification of the household position hierarchy gave point predictions for the age patterns of the shares that were very similar to those of the first specification, with two exceptions. The irregular age pattern of lone mothers aged 20-45 noted above for the first specification did not show up in the second one. On the other hand, the predicted age pattern for men who live with a cohabitee but without children (COH0) displays some unrealistic irregularities at ages around 50. This is caused by irregularities in the age patterns of estimated b(x) values for COH0, in particular for the ratios- Strong smoothing of the b(x) estimates in the age direction did not help.

Uncertainty results reported below apply to the first specification.

3.2 Standard deviations and correlations

Table 2 shows estimates of the standard deviations $\sigma_{\delta pk}$, $\sigma_{\theta pk}$, $\sigma_{\varepsilon pk}$, $\sigma_{\delta rk}$, $\sigma_{\theta rk}$, and $\sigma_{\varepsilon rk}$.

The first three columns contain standard deviations of the products, which determine the levels of the shares for men and women combined. The ratios in the last three columns relate to the differences between men and women. When we inspect the estimated standard deviations of the LC residuals ($\sigma_{\varepsilon pk}$ and $\sigma_{\varepsilon rk}$) we notice that (at least judged by this criterion) the levels of the shares are more difficult to predict than the differences between the sexes. The time series of COH0 as a share of COH0 and COH+ seems to be a process that is more difficult to model, both in terms of the LC model, and the Random Walk with Drift for the time index of the LC model, than the time series for the other cases. Residuals are large for ages over 65, in particular for the years 1998-2000 and in 2007. Adding an extra component $b_{2pk}(x).k_{2pk}(t)$ to the LC model in expression (11) might solve this problem. But similar to the case of cohort effects noted above, we have too few data points to obtain good estimates for such an extra component.

Table 2. Estimated standard deviations $\sigma_{\delta pk}$, $\sigma_{\theta pk}$, $\sigma_{\varepsilon pk}$, $\sigma_{\delta rk}$, $\sigma_{\theta rk}$, and $\sigma_{\varepsilon rk}$

	Products			Ratios		
k	$\sigma_{\delta pk}$	$\sigma_{\theta pk}$	$\sigma_{\varepsilon pk}$	$\sigma_{\delta rk}$	$\sigma_{\theta rk}$	$\sigma_{\varepsilon rk}$
1 (SIN0, MAR0 and MAR+)	3.28E-7	1.18E-6	0.1486	0.0016	0.0059	0.0127
2 (MAR0 and MAR+)	0.0505	0.1820	0.0693	0.0191	0.0688	0.0526
3 (MAR0)	0.0461	0.1663	0.6282	0.0774	0.2790	0.0309
4 (COH0 and COH+)	0.0295	0.1065	0.0309	0.0210	0.0759	0.0201
5 (COH0)	0.0901	0.3248	4.2808	0.1004	0.3623	0.0758
6 (CHLD)	0.0625	0.2252	1.8937	0.0019	0.0068	0.0577
7 (SIN+)	0.0776	0.2796	0.1281	0.0709	0.2555	0.0276
8 (INST)	0.0814	0.2936	0.1401	0.0238	0.0857	0.0198

Expression (22) gives the variance of the ξ-fractions in the logit scale. We encountered a problem for ξ_1, the logit transform of the total share of SIN0, MAR0, and MAR+. Estimated LC parameters $b_{p1}(x)$ for the products turned out be excessively large, with orders of magnitude of 1000 or larger (in absolute value). Estimates of $k_{p1}(t)$ were correspondingly small. Because of the squared value $b^2_{p1}(x)$ in the variance of ξ_1, the latter variance became excessively large. To solve this issue, we used an ad-hoc correction of the $b_{p1}(x)$, and made them equal to the average value of the $b_{p2}(x)$ and $b_{p3}(x)$ estimates.

The nature of this numerical problem is unknown. It is similar to unstable estimates one sometimes obtains for the parameters of a multivariate regression model when two or more independent variables are strongly correlated (multicolinearity). The second specification of household position hierarchy did not suffer from this problem.

The first order autocorrelation across ages in the LC residuals was estimated as 0.58 (median value across household positions), where we assumed that this correlation is the same for products and ratios, and that it is the same for each year. We also estimated the correlation between product residuals and ratio residuals and found a value equal to -0.02.

3.3 Prediction intervals for households

Table 3 gives predictions for private households of various types.

Table 3. Average value, coefficient of variation (CV), and lower and upper bounds of 67 per cent prediction intervals, for the number of private household, by household type. CV in per cent, other numbers in millions

	Men living alone	Women living alone	Married couple	Cohabiting couple	Lone father	Lone mother	All private households (incl. other)
2010: Observed	1.247	1.422	3.346	0.829	0.084	0.402	7.447
2020: Average	1.451	1.592	3.094	0.957	0.116	0.271	7.669
CV (%)	9.6	9.0	8.3	10.5	74.4	53.1	4.1
67% low	1.307	1.454	2.858	0.866	0.053	0.145	7.376
67% high	1.591	1.735	3.342	1.049	0.180	0.388	7.975
2030: Average	1.500	1.743	3.193	0.930	0.140	0.275	7.966
CV (%)	12.1	11.4	9.9	10.3	91.5	63.1	5.1
67% low	1.317	1.547	2.902	0.843	0.048	0.122	7.571
67% high	1.674	1.933	3.485	1.017	0.238	0.419	8.365
2040: Average	1.511	1.825	3.160	0.903	0.160	0.280	8.022
CV (%)	10.8	9.6	9.9	11.1	97.9	70.6	6.6
67% low	1.346	1.652	2.865	0.808	0.044	0.106	7.511
67% high	1.669	2.000	3.460	0.994	0.286	0.448	8.533

As noted above, predicted numbers of married or cohabiting men (with or without resident children) were not consistent with corresponding numbers for women, at least not initially. Numbers of households consisting of a married or cohabiting couple reported in Table 3 were computed as harmonic means of inconsistent numbers of men and women.

Predictions for households that are numerous are more certain than those for less numerous households, relatively speaking. By and large, Table 3 shows smallest coefficients of variation for married couple households, and largest values for lone father households. With a few exceptions, uncertainty as measured by the CV increases over time. Similar findings have been reported for probabilistic household forecasts for Norway, Denmark, and Finland,

although these were computed using extrapolation methods that are very different from ours (Alho and Keilman 2010; Christiansen and Keilman 2012).

How do the results in Table 3 compare with the probabilistic household forecast of Statistics Netherlands? Point predictions and 67 per cent prediction intervals for selected household positions and household types are available (in Dutch only) at http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=80987NED&D1=3-25&D2=0,3-4&D3=9,19,29&HDR=T&STB=G1,G2&VW=T . For the year 2040, Statistics Netherlands (SN) predicts 8.478 million private households, with a 67 per cent interval of [7.839, 9.067]. Our prediction is slightly lower, a bit more certain, but it falls well within the SN interval. There is also close agreement between our results and those of Statistics Netherlands for household types "Couple" (married or cohabiting) where SN predicts 4.239 million [3.734, 4.709], "Men living alone" 1.712 [1.321, 2.088], and "Women living alone" 1.912 [1.544, 2.261]. However, our household forecast predicts much fewer lone mothers. The SN prediction of lone mothers in 2040 is 0.466 million. This number falls outside our 67 per cent prediction interval, and the SN prediction is much more in line with historical values than ours, which shows a strong fall in the beginning of the forecast period; see Table 3. Figure 8 shows that our share predictions for lone mothers are much lower than those of SN for ages 40-64 in particular, because our method is not able to take account of a cohort effect in the shares. Note, at the same time, that our predictions for lone parents are extremely uncertain, with CV values in 2040 of 71 (lone mothers) and 98 (lone fathers) per cent. These CV values are one order of magnitude larger than those for other household types.

As noted above, the prediction for all private households in 2040 by Statistics Netherlands is more uncertain than our prediction. It is difficult to assess the reason why this is the case, because correlations in the shares across household positions and across ages play an important role here. These correlations lead to correlated numbers of households of various types, which could be assessed empirically. We have not done this, because we do not have the necessary information from the SN simulations. Nonetheless we note that prediction intervals for couples, men who live alone and women who live alone are wider for the predictions of Statistics Netherlands than for our predictions. SN predictions for lone fathers and lone mothers are much more certain than ours, but these have little weight in the predictions for the overall numbers of private households.

4. Conclusions

We have computed a stochastic household forecast for the Netherlands for the period 2010-2040 by the random share method. Time series of shares of persons in nine household positions, broken down by sex and five-year age group for the years 1996-2010 were modelled by means of the Hyndman-Booth-Yasmeen product-ratio variant of the Lee-Carter model. This approach is able to preserve age patterns and differences between men and women. We modelled the two time indices for each household position as a Random Walk with Drift (RWD), and computed prediction intervals for them, taking into account the correlation between ages. This gave us prediction intervals for random shares. The latter

shares were combined with population numbers from an independently computed stochastic population forecast of the Netherlands.

By and large, predicted age patterns of the shares for nine household positions looked reasonable, with a few important exceptions. Unrealistic patterns arise when the observed time series contain cohort effects. In our data set this was the case for cohabiting persons, persons who live alone, and lone parents. The Lee Carter model as we have used it accounts for age patterns and period progression in the data, but it does not model cohort progression and cohort effects. In principle one could include the latter type of effects in the Lee Carter model, by adding an extra component. However, one needs a long time series of data to identify such a cohort component. In addition, one has to overcome the technical problem of identifying separate effects for age, period, and cohort. Since we have only 15 years of data, we have not done this. Instead we applied ad-hoc solutions to the problem of unrealistic age patterns for some household positions.

Our simulations results are very similar to the probabilistic household forecast computed by Statistics Netherlands (SN), with some exceptions. The SN point predictions are obtained by means of a multistate cohort component model, which is able to account for cohort effects. Therefore we have more confidence in the SN point predictions than in ours. Overall, our predictions are a bit more precise, with the exception of predictions of households for lone fathers and lone mothers. Uncertainty in the household-shares in the SN forecast is derived from intuitively chosen prediction intervals, similar to the practice of choosing high or low fertility, mortality and migration assumptions when deriving variants of population projections. Moreover, SN assumed perfect correlations in the forecast errors of the shares across ages and in the time dimension. Our uncertainty assessments are empirically based, hence our method is more transparent and more objectie than that of Statistics Netherlands. To justify the more conservative intervals used by Statistics Netherlands, arguments should be given why the uncertainty in the coming decades is substantially higher than that contained in the time-series since 1995.

The general conclusion of this paper is that our method is useful for generating errors around expected values of shares that are computed independently. We have used the residual errors of our fitted Lee Carter model as a basis for our prediction intervals. As stated above, our method does not account of cohort effects in the data. Thus the fit of the Lee Carter model might be improved and the residual errors might be reduced if one could include such effects (provided the necessary data are available). For that reason, we consider our uncertainty assessments as conservative, in the sense that they resulted in relatively wide prediction intervals.

References

Alders, M. (1999). Stochastische huishoudensprognose 1998-2050 (“Stochastic household forecast 1998-2050”). Maandstatistiek van de Bevolking 47 (11): 25–34

Alders, M. (2001). Huishoudensprognose 2000-2050: veronderstellingen over onzekerheidsmarges (“Household forecast 2000-2050: assumptions on uncertainty intervals”). Maandstatistiek van de Bevolking 49 (8): 14–17

Alho, J. and Keilman, N. (2010). On future household structure. Journal of the Royal Statistical Society Series A 173(1): 117-143

Carolina, N. and Van Duin, C. (2010) Onzekerheidsmarges voor de sterfteprognose van het CBS (“Uncertainty intervals for the mortality forecast of the CBS”). Bevolkingstrends 58 (2): 32-37

De Beer, J. and Alders, M. (1999). Probabilistic population and household forecasts for the Netherlands. Joint Economic Commission for Europe–EUROSTAT Work Session on Demographic Projections, Perugia, May 3rd–7th. Geneva: Economic Commission for Europe. (Working paper 45)

Christiansen, S. and Keilman, N. (2012). Probabilistic household forecasts based on register data -the case of Denmark and Finland (submitted)

Hyndman, R., Booth, H. and Yasmeen, F. (2012). Coherent mortality forecasting: The product-ratio method with functional time series models. Demography DOI: 10.1007/s13524-012-0145-5 (published online 10 October 2012)

Van Duin, C. and Harmsen, C. (2009). Een nieuw model voor de CBS huishoudensprognose (”A new model for the CBS household forecasts”). Bevolkingstrends 57 (3): 20-42.

Van Duin, C. and Stoeldraijer, L. (2011). Huishoudensprognose 2011–2060: meer en kleinere huishoudens (“Household forecast 2011-2060: more and smaller households”). Bevolkingstrends 59 (4): 59-67.

Figure 1. Point predictions of age-specific shares for women in household position MAR+. Full RWD extrapolation.

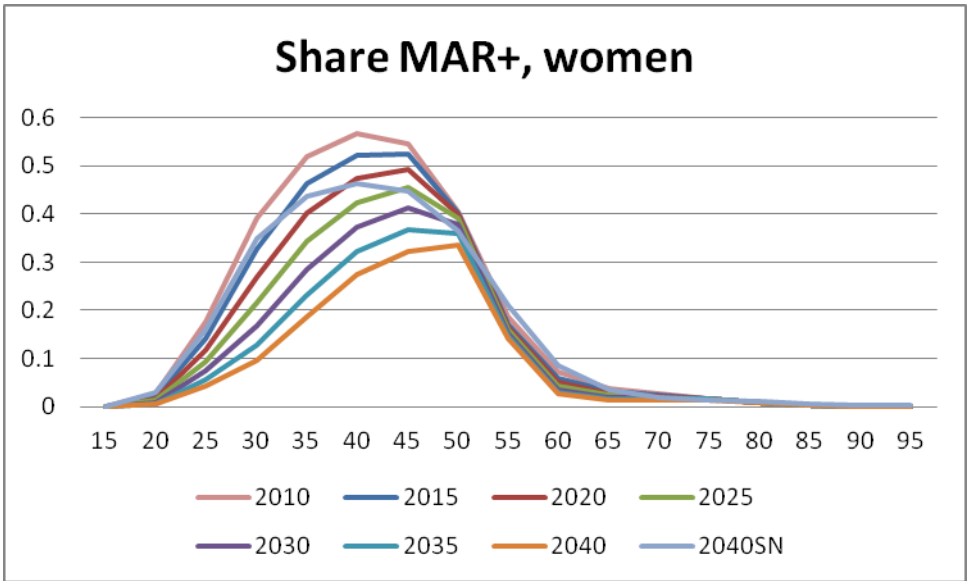


Figure 2. Point predictions of age-specific shares for women in household position COH+. Full RWD extrapolation

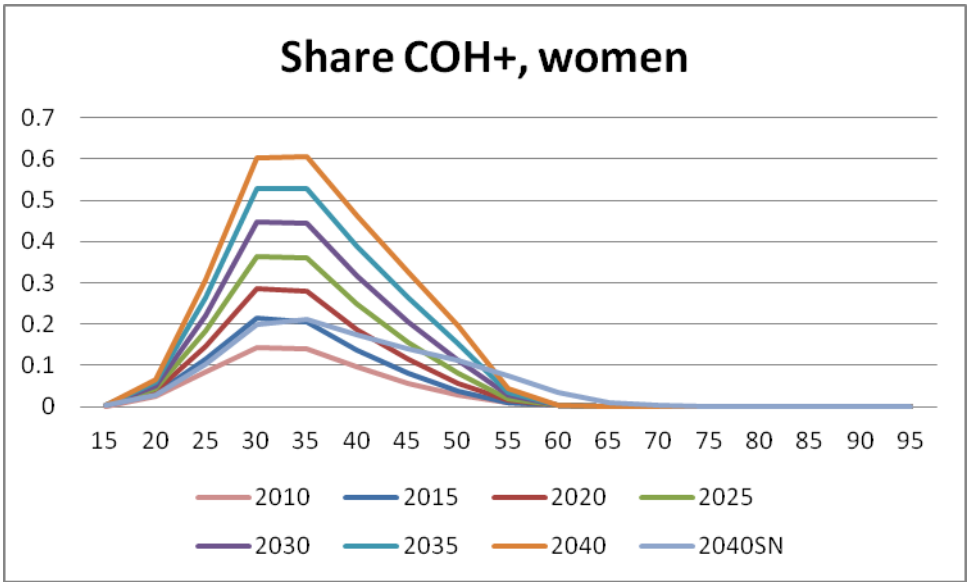
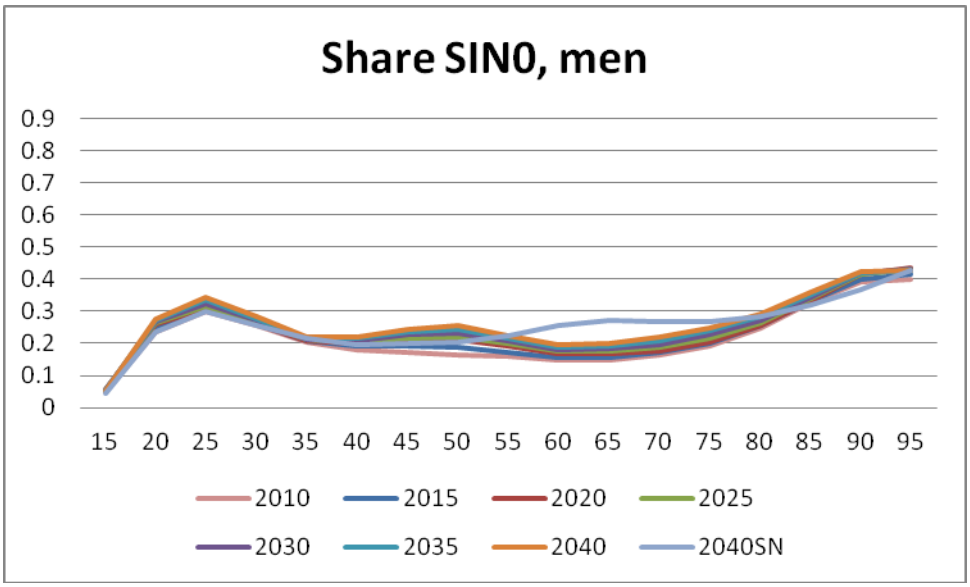


Figure 3. Point predictions of age-specific shares for men and women in household position SIN0



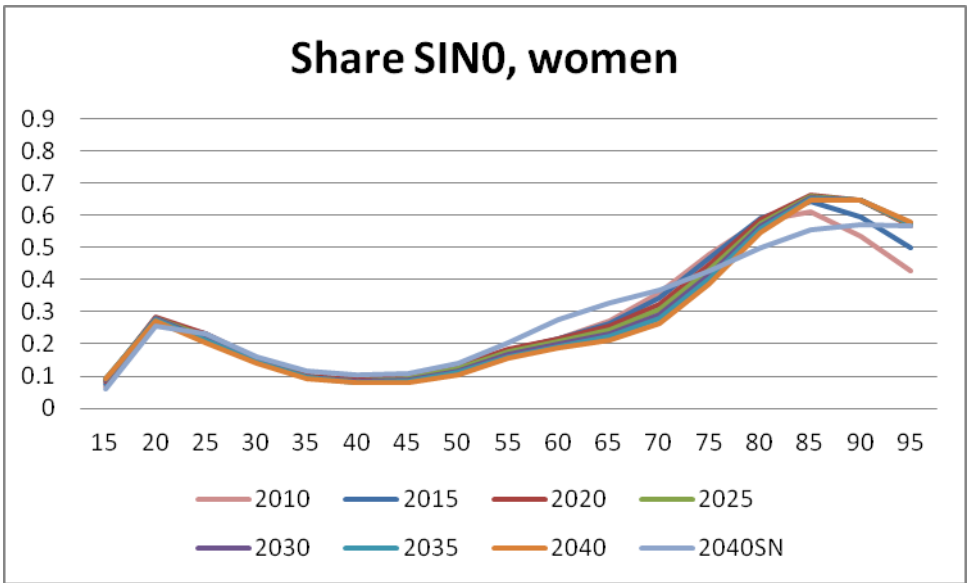
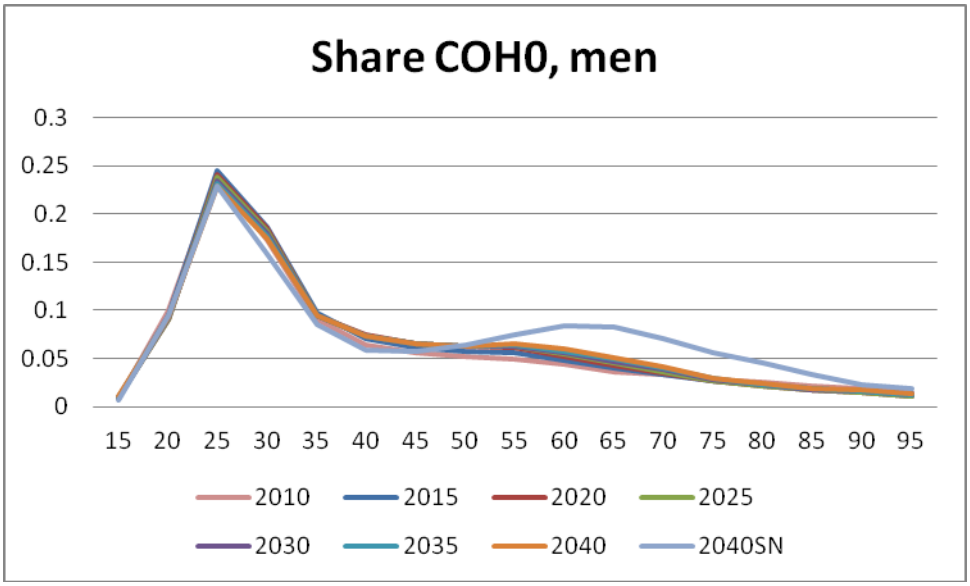


Figure 4. Point predictions of age-specific shares for men and women in household position COH0



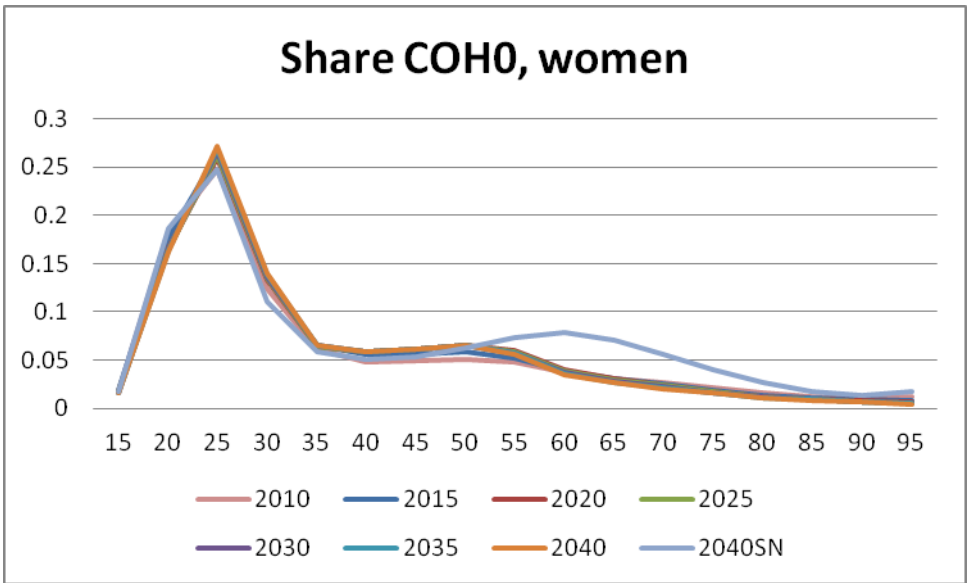
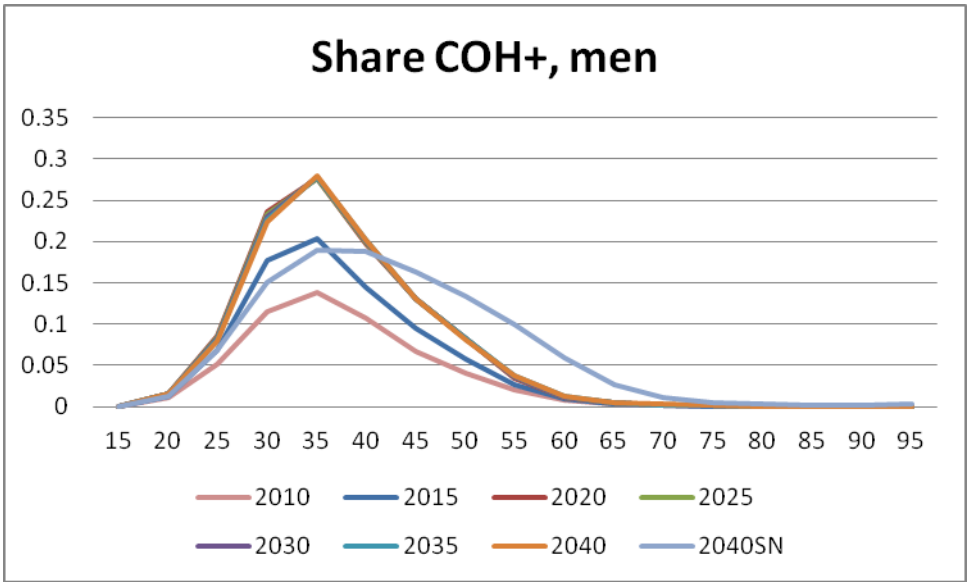


Figure 5. Point predictions of age-specific shares for men and women in household position COH+



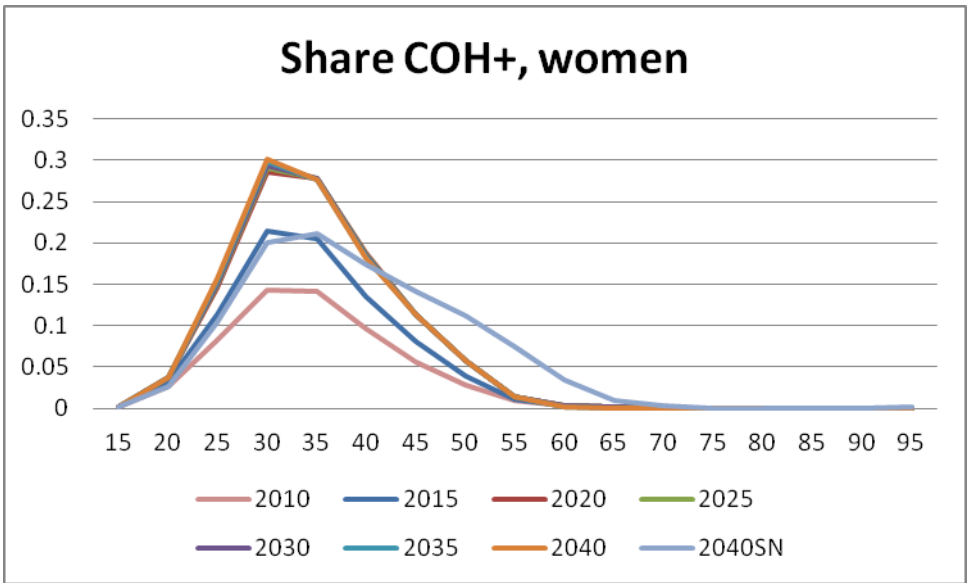
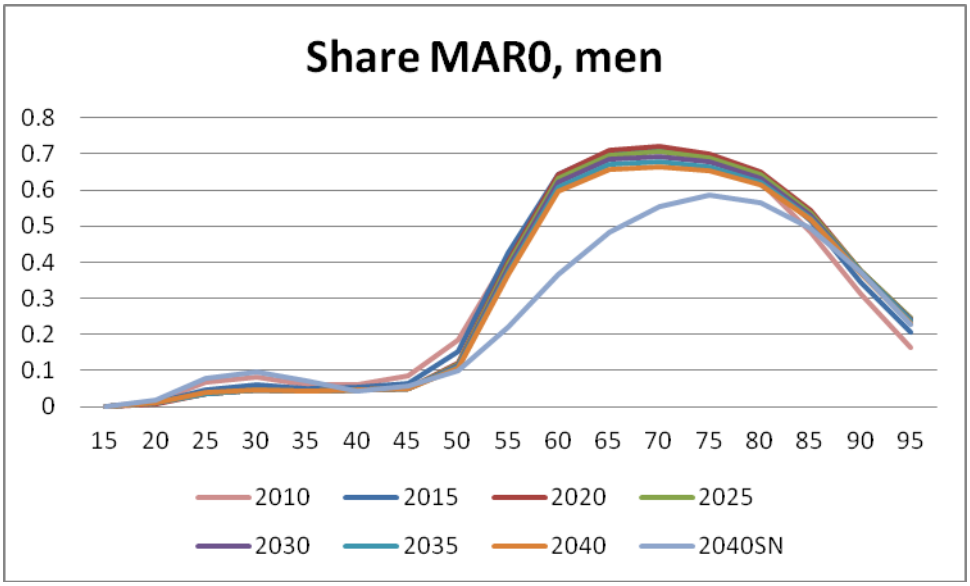


Figure 6. Point predictions of age-specific shares for men and women in household position MAR0



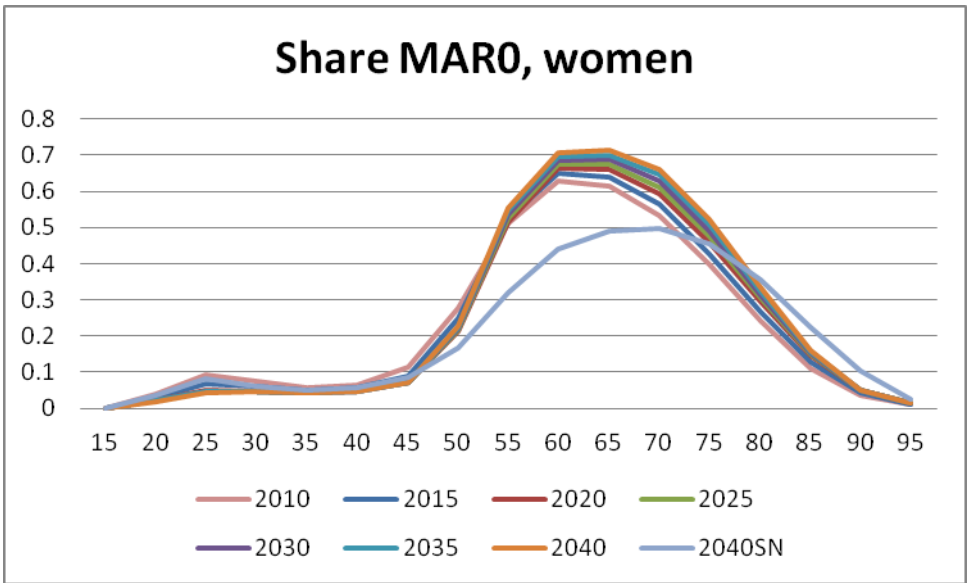
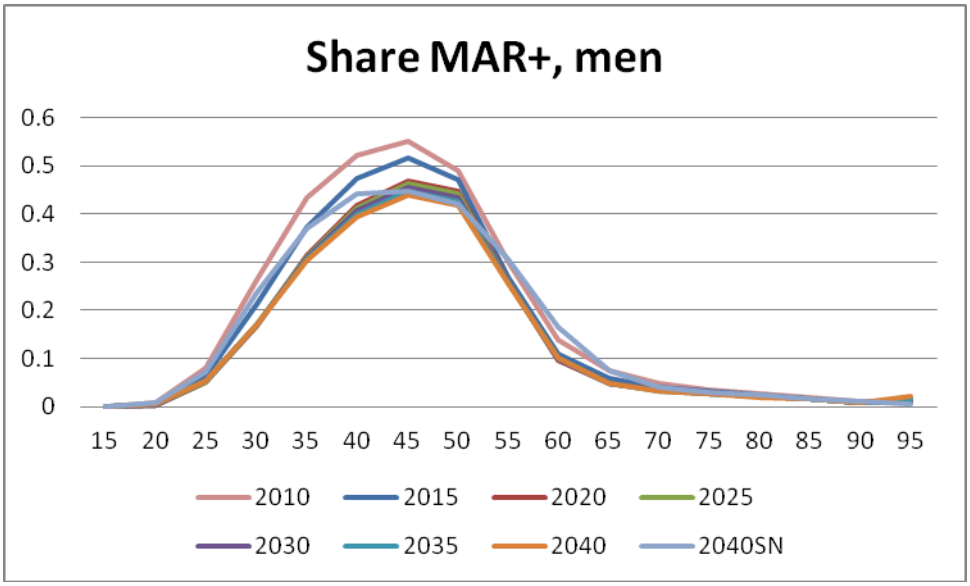


Figure 7. Point predictions of age-specific shares for men and women in household position MAR+



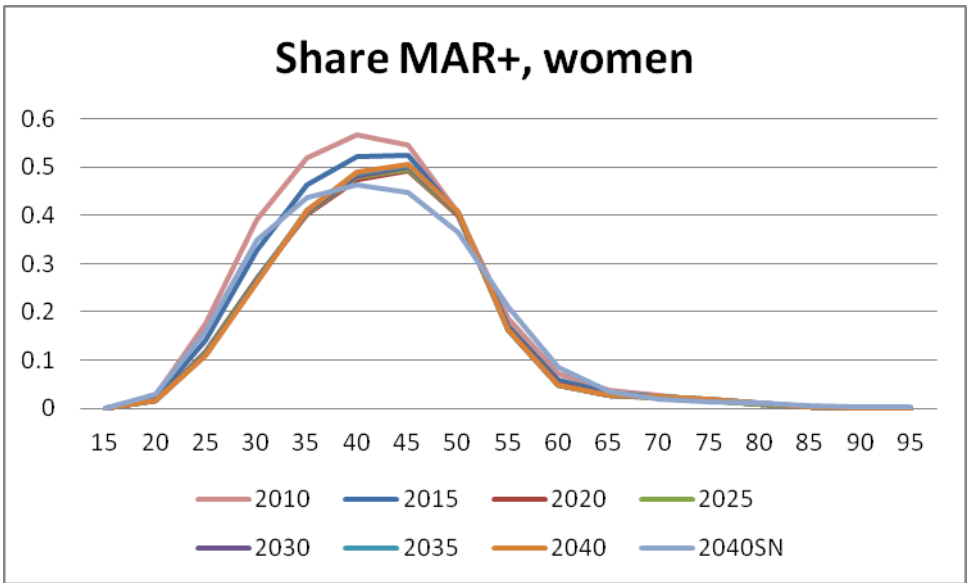
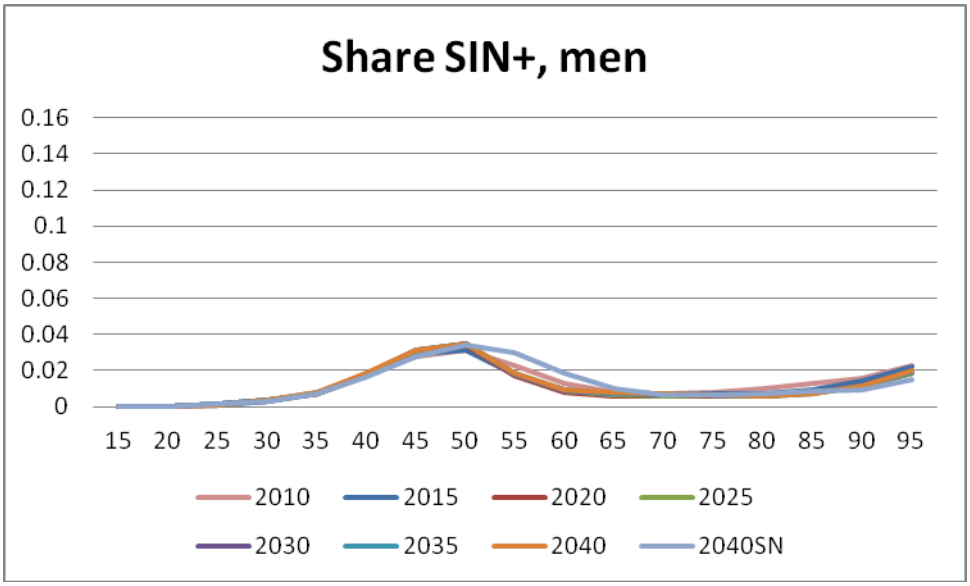


Figure 8. Point predictions of age-specific shares for men and women in household position SIN+



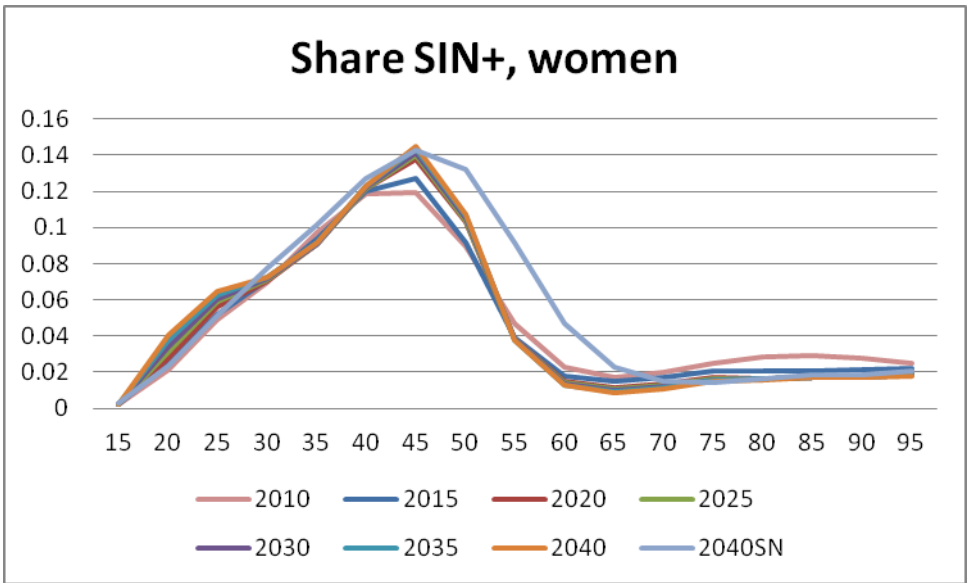


Figure 9. Point predictions of age-specific shares for men and women in household position INST