Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2010)
173, *Part* 1, *pp.* 117–143

On future household structure

Juha Alho

University of Joensuu, Finland

and Nico Keilman

University of Oslo, Norway

[Received December 2007. Final revision March 2009]

Summary. We develop a method for computing probabilistic household forecasts which quantifies uncertainty in the future number of households of various types in a country. A probabilistic household forecast helps policy makers, planners and other forecast users in the fields of housing, energy, social security etc. in taking appropriate decisions, because some household variables are more uncertain than others. Deterministic forecasts traditionally do not quantify uncertainty. We apply the method to data from Norway. We find that predictions of future numbers of married couples, cohabiting couples and one-person households are more certain than those of lone parents and other private households. Our method builds on an existing method for computing probabilistic population forecasts, combining such a forecast with a random breakdown of the population according to household position (single, cohabiting, living with a spouse, living alone etc.). In this application, uncertainty in the total numbers of households of different types derives primarily from random shares, rather than uncertain future population size. A similar method could be applied to obtain probabilistic forecasts for other divisions of the population, such as household size, health or disability status, region of residence and labour market status.

Keywords: Household forecast; Norway; Population forecast; Probabilistic forecast; Random shares

1. Introduction

Knowledge about the future number of households, and their composition, is important for policy purposes. The amount of support that is needed by the elderly depends on the numbers living alone, for example (Grundy, 2001; Glaser *et al.*, 2003). Housing planners use forecasts of household size to anticipate changes in the demand of dwellings with elevators and wheelchair access (Holmberg, 1987; King, 1999; Muller *et al.*, 1999). A recent environmental concern stems from the increasing number of small households. Household members share space, home furnishings, transportation and energy, leading to significant economies of scale. For instance, members of two-person households in the USA in 1993–1994 used 17% less energy per person than one-person households did (O'Neill and Chen, 2002). Thus, even when the population remains constant, more small households imply a larger demand for resources.

There are many possible future household developments for a given population, but some of these are more likely than others. As opposed to a deterministic forecast, which predicts *only one number* (or perhaps just a few; see below) for a certain year, a probabilistic forecast tells us how likely it is that future household numbers will be within a certain *range*. For instance, the results that we obtain in Section 4 imply a 50% chance that the number of private households in

Address for correspondence: Nico Keilman, Department of Economics, University of Oslo, PO Box 1095, Blindern, Oslo, N-0317, Norway.
E-mail: nico.keilman@econ.uio.no

 0964–1998/10/173117

Norway will be between 2.51 million and 2.69 million in 2030, up from an observed 2.03 million in 2002. Information of this kind allows policy makers, planners and other forecast users in the fields of housing, energy, social security etc. to take appropriate decisions, because some household variables are more uncertain than others. It also guides them once actual developments start to deviate from the most likely path. New actions or updated plans are unnecessary as long as developments remain close to the predicted path. Deterministic forecasts traditionally deal with forecast uncertainty by formulating alternative scenarios, usually in terms of a high and a low trajectory for some key input parameter, in addition to a most likely trajectory (Jiang and O'Neill, 2006). The drawback is that uncertainty is not quantified, and hence the user does not know how likely it is that the high trajectory will materialize, instead of the most likely trajectory. Moreover, the results are not plausible from a statistical point of view, as they implicitly assume perfect correlation across age, time and type of household (Lee, 1999; Alho *et al.*, 2008).

The purpose of this paper is to show how statistical forecasting methods and existing software can be used to produce a probabilistic household forecast, thus avoiding the drawbacks of a deterministic approach. Methods for computing *probabilistic population* forecasts have become well established in the past two decades. The tradition goes back to Törnqvist (1949), who probably was the first to integrate probabilistic thinking in population forecasting. Examples include probabilistic population forecasts for the USA, Austria, Finland, the Netherlands and Norway by Lee and Tuljapurkar (1994), Lutz and Scherbov (1998), Alho (1998), De Beer and Alders (1999) and Keilman *et al.* (2002) respectively, for 18 European countries (the old European Union 15 countries plus Norway, Iceland and Switzerland) by Alho *et al.* (2006), for major world regions by Lutz *et al.* (1996, 2001) and for all countries in the world in National Research Council (2000). *Deterministic household* forecasts have a somewhat longer tradition (US National Resources Planning Committee, 1938; United Nations, 1973; Keilman *et al.*, 1988), but few scholars have attempted to compute a *probabilistic household* forecast.

Alders (1999, 2001) combined a probabilistic population forecast with random shares that distributed the population probabilistically over six household positions: individuals could either 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, Alders computed a predictive distribution for the number of lone mothers aged 40 years in 2015 as the product of two random variables, namely the number of women aged 40 years in 2015 and the share of those women who live as a lone mother. Expected values for population variables and shares for specific household types were obtained from observed time series, but uncertainty distributions that were assumed for both were based on intuitive reasoning, and correlations across age, sex and time were disregarded. Scherbov and Ediev (2007) combined a probabilistic population forecast for the population broken down by age and sex with random headship rates. In demography, a headship rate reflects the proportion of the population that is the head of the household, for a given combination of age and sex (United Nations, 1973; Jiang and O'Neill, 2004). Like Alders, Scherbov and Ediev based a large part of their uncertainty distributions on intuition. In contrast, our contribution is to show how uncertainty in the forecast of the shares that distribute the population over several household positions can be modelled as a stochastic process, the parameters of which can be estimated from past data.

We discuss conceptual issues in Section 2. Perhaps surprisingly, there is no general agreement on the concepts of household and family. We give our working definitions and discuss measurement problems. In Section 3, we briefly describe deterministic and probabilistic forecasts of households and populations, and present our model that is based on random shares. Section 4 gives an empirical application for Norway.

2. Conceptual issues

2.1. Defining household and family

Durkheim's classic notion of a 'nuclear family' refers to a married couple and their dependent children, living together but apart from other kin (Moen and Forest, 1999; Hoffmann-Nowotny, 1987; Smith, 1968). During the post World War II period, the notion of a family has broadened to include a range of forms, but new debates have arisen (Settles, 1999). For instance, should a cohabiting or a same-sex couple be considered as a family? Does the family definition also cover lone parents? Such issues are important when families are viewed from the perspective of the special rights and obligations that family members have with each other (Lenoir, 2007).

As our interest is in the sharing of dwellings, we shall bypass the thorny issues that were referred to above and consider a less value-laden concept of a household. A *household* consists of all people who live in the same dwelling, independently of how they view their family relationships with each other. Thus, a household may include just one person, or several unrelated individuals (e.g. students). These definitions are compatible with the so-called 'dwelling definition' of the Conference of European Statisticians (2006). With this approach we could logically take dwellings, or households, as our statistical units of observations and analyse how the number of households of different types (for instance one-person households, couple households and one-parent family households) develop over time. However, opting for the household as the unit of analysis would lead to conceptual problems. The moment at which a household is formed or dissolved is not always well defined; nor is the household membership always clear. Rules may be set up, for instance that two households, when observed at two different points in time, are the same as soon as they have the same head of household. But rules of this kind are largely arbitrary and, moreover, may have a strong influence on the results (McMillan and Herriott, 1985; Keilman and Keyfitz, 1988). Data on an individual person, such as date of birth, death, marriage, divorce or separation and leaving home, provide a firmer basis, once the individuals' household positions have been defined. This is not without problems either, but the issues appear to be solvable. We shall now address these.

A fairly generally recommended practice is to distinguish the following positions in a private household (Moen and Forrest, 1999).

(a) *Child* refers to a blood, stepson or stepdaughter, or adopted son or daughter who is younger than 25 years of age (but regardless of marital status) who lives in the household of at least one of the parents and who has no partner or own child(ren) in the same household.
(b) A *cohabiting couple* (i.e. partners living in a consensual union) is a couple that has a marriage-like relationship, irrespective of the partners' sexes. Cohabiting people can have any marital status (including married; in that case they are legally married to different partners, who live elsewhere).
(c) The category of *married couples* consists of those who are currently married and live together with the spouse.
(d) A *lone parent* lives together with one or more children as defined above, but without a spouse or cohabiting partner.
(e) An adult may *live alone*.
(f) Young adults aged 25 years or over, as well as people who are younger than 25 years with a partner or an own child, who live with their parent(s) belong to the category *other*. This last category also includes people who live in a multiperson household but who have no relationship (parent–child, or partner in consensual or marital union) to the other household members.

Knowing the household positions of the household members, the type of household can be deduced: a private household may be a married couple household, a cohabiting couple household, a lone parent household, a one-person household or a household of type 'other'. Compared with Alders's household forecast, our forecast distinguishes married couples from cohabiting couples, but (owing to limitations of the data) we ignore institutional households.

2.2. Some measurement problems

It is not always possible to elicit reliable information on household position in surveys or censuses. We point out four recognized sources of potential variation.

(a) The notion of 'living in the same dwelling' is based on an individual's place of usual residence. In our Norwegian application we shall use data from the census and the Survey of Living Conditions (SLC). The SLC is a panel survey with annual waves in the spring of each year since 1997; see Normann (2004). The census is based on the *de jure* definition, whereas the SLC uses a *de facto* concept that includes the requirement of 'common housekeeping'. These differ, in particular, for young adults. For example, a person who leaves the parental household will not be officially registered at the new address, unless he or she marries, has a child or receives the major part of his or her income from labour. Similarly, when elderly people move back and forth between a private dwelling and an institution because of health problems, it is not clear how their place of usual residence is decided.

(b) The distinction between private and institutional households is not clear cut in intermediate forms such as assisted living.

(c) The notion of a cohabiting couple living in a 'marriage-like relationship' leaves room for subjective interpretations, which may change over time and differ between the partners.

(d) A lone parent who starts cohabiting may still be classified as a lone parent, if the new partner does not take parental responsibility for the children.

3. Method

3.1. Outline

We start with a non-technical outline of our approach, whereas Sections 3.2–3.5 contain formal details. We build on an existing probabilistic forecast for the population of Norway. Similar to a deterministic population forecast, a probabilistic population forecast is based on the 'book-keeping' identity

$$\text{population}(t+1) = \text{population}(t) + \text{births}(\text{during } t) - \text{deaths}(\text{during } t) \\ + \text{immigration}(\text{during } t) - \text{emigration}(\text{during } t).$$

In practice, updates of this kind are done for men and women by age. In a deterministic population forecast, we need to specify age-specific fertility rates for women and mortality rates for women and men, for all future years of interest. These are used to calculate numbers of births and deaths. Migration is usually specified as net migration in absolute numbers, the survivors of which are added to the population at time t. Starting from a known population by age and sex at some initial time $t=0$—the so-called jump-off population—this method is used recursively to keep track of the resulting changes in population, by age and sex. Stochastic (or probabilistic) population forecasts are similar, but in this case future fertility and mortality rates and net migration numbers are considered as random variables (e.g. Alho and Spencer (2005)). Simulation is used to carry out the propagation of uncertainty, from the rates to future population numbers. A joint distribution that is derived in this manner for the future demographic variables can be called a predictive distribution. In other words, it is the probability distribution of the future demographic variables, conditional on what was known at the time when the forecast was made.

For the 18 European countries that were mentioned in Section 1, the results of such probabilistic population forecasts are available as simulated predictive distributions for the numbers of men and women by 5-year age group for the period 2010–2050, at 10-year intervals. The predictive distribution of each population variable is given in the form of 3000 simulated values; see `www.stat.fi/tup/euupe/index_en.html`. We use simulation results for Norway from this source. Next, given age, sex and time, we assume that the population will have one of the six mutually exclusive household positions that were defined in Section 2.1, and we define a set of six shares, the sum of which equals 1. We predict the shares as a function of age, sex and time. This requires specifying their predictive distributions. We assume a particular statistical model for each share, which defines its distribution at one point in time, as well as its development over time. The parameters of the model are the expected value of the share and its variance. We derive the shares' expected values from a deterministic household forecast. The variance of each share reflects the deviation from its expected values; in other words, the variance is a measure for forecast errors in the shares. Empirical information for forecast errors is derived from the errors of a household forecast, prepared in the early 1990s and covering the years 1990–2020. We assume that the errors in the future are not smaller. Finally, we compute a forecast for the number of people in a given household position as the product of two random variables, namely the predicted population and the predicted share for that household position.

In practice, we draw 3000 values from the predictive distribution of the shares for each combination of age, sex and household position, for the years 2010, 2020 and 2030, and combine them with 3000 simulated values for the population numbers by age and sex. Given predicted numbers of *people* in all six household positions in a certain year, predicted numbers of *households* of various types can be derived.

3.2. Notation

We shall write $V(j,x,s,t)$ for the number of people in household position $j=1,2,\ldots,6$ who are in age $x=0,1,\ldots,\omega$, sex $s=1$ for males and $s=2$ for females, at time $t=0,1,2,\ldots$. Aggregating over position, we obtain the population $W(x,s,t)=\Sigma_j V(j,x,s,t)$ of age x and sex s at time t. The share of household position j is $\rho(j,x,s,t)=V(j,x,s,t)/W(x,s,t)$. We have a probabilistic forecast for the population $W(x,s,t)$ available. We shall specify a predictive distribution for the shares $\rho(j,x,s,t)$ for future years. Then, the product $\rho(j,x,s,t)\ W(x,s,t)$ will yield a predictive distribution for $V(j,x,s,t)$.

3.3. Stochastic forecast of population

The stochastic population forecast of Norway (the random variable $W(x,s,t)$) was taken from the project 'Uncertain population of Europe', which gives predictive distributions of the population by sex and age in 18 European countries for the period 2004–2050. A general description of the results is available in Alho *et al.* (2006), whereas Alders *et al.* (2007) and Alho *et al.* (2008) discuss the assumptions that are made. Simulation results are posted at `www.stat.fi/tup/euupe/no11_results_nor.html`. For Norway, a population of 5.26 million is expected for 2030, which is very close to the 5.37 million of Statistics Norway's medium growth variant, made in 2005 (`www.ssb.no/emner/02/03/folkfram/tab-2005-12-15-01.html`). The

80% prediction interval is [4.90, 5.65] million. An 80% interval for the share of the population aged 65 years and over is [19.5, 23.3]%— in 2008 it was 14.7%.

3.4. Point predictions of household shares

The point predictions (or expected values) of the household shares were derived from an updated deterministic household forecast for Norway (see Keilman and Brunborg (1995) for the original household forecast). This resulted in point predictions $\hat{\rho}(j, x, s, t)$ for the period 2002–2032. We used the program LIPRO ('Lifestyle projections') version 4.0 to compute the household forecast. LIPRO (see `www.nidi.knaw.nl/en/projects/270101/` and Van Imhoff and Keilman (1991)) is based on the methodology of multistate demography but includes several extensions to solve the particular problems of household modelling.

3.4.1. Population structure

The number and types of household position that we selected are a compromise between conflicting arguments. On the one hand, we want to have many household positions, to provide the user of the forecast with detailed results. But, on the other hand, the available data restrict the possibilities that we have to a considerable extent. As opposed to the random shares, for which we distinguish only six household positions (see Section 2.1) we had enough data for the deterministic household forecast to work with a slightly more detailed classification. We have used the following household positions for individuals (positions 1–9 respectively):

(a) dependent child living with one or both parents;
(b) living in a consensual union without dependent children;
(c) living in a consensual union with dependent children;
(d) living with a spouse without dependent children;
(e) living with a spouse and dependent children;
(f) a person living in a one-person household;
(g) a single mother or father;
(h) living in a private household, but not in any of the positions (a)–(g);
(i) living in an institution for the elderly.

The institutionalized population was restricted to ages 65 years and older. Compared with the six household positions for which we shall specify uncertainty distributions, the nine positions that are listed here imply full coverage of the population, both in private and in institutional households. Also, we distinguish cohabiting couples and married couples with and without dependent children in their homes. The reason is that childless couples have a different childbearing behaviour from that of those with one or more children. Given the point predictions for the nine shares, those for the six shares are easily obtained by combining those for cohabiting or married people with and without children, and conditioning on being a member of a private household.

LIPRO implements a multistate demographic model that focuses on flows between the states. Multistate projections are inherently more complex than single-state projections. In particular, they require more extensive data. We shall here sketch the main aspects of the LIPRO calculations and refer to Van Imhoff and Keilman (1991) for the full details.

Let $\mathbf{V}(t)$ be a row vector of population counts at exact time t. We use 5-year age groups 0–4, 5–9, ..., 85–89 and 90 years and older. Their number is 19, and we shall write $\mathbf{V}(t) = (\mathbf{V}(0, t), \ldots, \mathbf{V}(18, t))$. We keep track of females and males separately, so $\mathbf{V}(x, t) = (\mathbf{V}(x, \mathrm{F}, t), \mathbf{V}(x, \mathrm{M}, t))$ for each age group $x = 0, \ldots, 18$. There are nine household positions, so, for example,

$\mathbf{V}(x, \mathrm{F}, t) = (V(1, x, \mathrm{F}, t), \ldots, V(9, \mathrm{F}, x, t))$, where $V(j, x, \mathrm{F}, t)$ is the number of females in position j, in age x, at time t. Thus, $\mathbf{V}(t)$ has a total of $19 \times 2 \times 9 = 342$ elements.

To describe population change, suppose first that there is no external migration. Consonant with age grouping, we take a unit of time to correspond to 5 calendar years. Consider survival from exact time t to exact time $t+1$ for women. This can be expressed as $\mathbf{V}(x+1, \mathrm{F}, t+1) = \mathbf{V}(x, \mathrm{F}, t)\, \mathbf{S}(x, \mathrm{F}, t)$, where $\mathbf{S}(x, \mathrm{F}, t)$ is a 9×9 matrix of origin–destination-specific survival probabilities for women from age group x to age group $x+1$. Specifically, row j of $\mathbf{S}(x, \mathrm{F}, t)$ gives the proportions surviving to positions $k = 1, \ldots, 9$, for women who are in position j at time t. Since the proportions reflect both mortality and position transitions, the row sums are typically less than 1. For forecasting purposes the probabilities in $\mathbf{S}(x, \mathrm{F}, t)$ could be specified directly but, as empirical data often come in the form of occurrence–exposure rates (which are equivalent to hazard rates in survival analysis) rather than proportions, LIPRO expects as inputs age and sex-specific occurrence–exposure rates from one household position to another. It is well known (e.g. Schoen (1988)) that if the rates are equal in all ages belonging to age group x, and they remain constant from exact time t to exact time $t+1$, then $\mathbf{S}(x, \mathrm{F}, t)$ is given by an exponential function of a matrix argument. We skip the details but note that the argument matrix is a function of the occurrence–exposure rates that are specific for age, sex and household position, including the death rates. Survival calculations in LIPRO are based on this assumption of time constant rates, which, of course, is an approximation to how things are in reality (for a discussion of alternatives, see Alho and Spencer (2005)).

Consider births. As is usual in demography, LIPRO attributes births to women only. The child-bearing ages in our application are 15–49 years, corresponding to $x = 3, \ldots, 9$. LIPRO expects a vector of age and household-position-specific (annual) birth rates $\mathbf{b}(x, t)$ that are assumed to be the same for all women in an age group. Being analytical, the exponential solution to survival calculations gives us not only $\mathbf{V}(x+1, \mathrm{F}, t+1)$ but also $\mathbf{V}(x+t', \mathrm{F}, t+t')$ for any $0 < t' < 1$. These values can be analytically integrated over t' to obtain a vector with person-years by age and household position, $\mathbf{K}(x, \mathrm{F}, t)$, during the period $(t, t+1)$. LIPRO uses this formulation. Multiplying $\mathbf{K}(x, \mathrm{F}, t)$ elementwise by $\mathbf{b}(x, t)$ yields a vector of births, by household position, in age group x, that are expected during the period from t to $t+1$. LIPRO survives these to exact time $t+1$ by using an exponential survivorship matrix of the type that was discussed above. Only now survival time is 0.5 time units, corresponding to 2.5 calendar years. This is an approximation that is frequently used in demography.

The description of external migration is logically difficult, because in-migration does not have a natural 'population at risk'. In this respect there is no difference between multistate and single-state projections. In applications, in which the level of migration is small relative to the populations of interest, a practical expedient is to express both in-migration and out-migration in terms of absolute numbers. But, as soon as this decision has been made, it becomes clear that it is not necessary to keep track of in- and out-migration separately. To see this, consider a woman entering household position j, in age group x, at some time $t+t'$, $0 < t' < 1$. We can think of her as starting a subpopulation of her own that survives to different household positions and renews itself by births, until $t+1$ and beyond, as described above. However, if a similar woman should leave the population of interest at $t+t'$, we should obviously subtract the subpopulation that she contributes after $t+t'$. From these considerations it is clear that it is simply the difference of in-migration and out-migration at any given time that determines the effect of external migration on the future evolution of the population. LIPRO expects as input age, sex and position-specific net migration numbers that are assumed to be uniformly distributed over the interval $(t, t+1)$, and it computes their net effect on the basis of the principles laid out above for survival and births.

The model is driven forward by assumptions, for each time interval, on net immigration by age, sex and household position, on occurrence–exposure rates for household events and mortality, and on birth rates; see below. To specify the jump-off population ($\mathbf{V}_0$) we used data from the Population and Housing Census of Norway, November 3rd, 2001, and assumed that the data would apply to January 1st, 2002. Readers who are interested in the census may consult Statistics Norway's Web site on this topic (in English) at `www.ssb.no/english/subjects/02/01/fob2001_en/`.

3.4.2. Events

For the events, consider first the eight positions for private households. There are $7 \times 8 = 56$ possibilities for changes of position, but not all are logically possible. For instance, a person cannot change directly from 'cohabiting, no children' to 'married, with children'. We estimated occurrence–exposure rates for changes of private household position from the SLC by using data from 1997–2002, for ages 16–79 years. See Normann (2004) for a technical description of the SLC. In 1997, the sample size was approximately 5000. We deduced household events from changes in reported positions in consecutive interviews. We discarded information about two or more household events in one calendar year. This occurred rarely. Similarly, owing to both left and right censoring, a small number of events could not be uniquely classified. Added over all five calendar years, we obtained 3645 events for 27 types of positional change, and 22462 years of exposure. The occurrence–exposure rates are specific for 5-year age groups, private household position and sex. Data on entries into and exits from institutional households are extremely scarce. We have used the entry rates and exit rates that were used in the previous household forecast of Keilman and Brunborg (1995).

Birth rates by 5-year age group and household position of the mother were taken from the previous household forecast and adjusted proportionally to match the numbers that Statistics Norway had registered for 2002–2006. *Death rates* by 5-year age group, sex and household position were estimated on the basis of deaths and exposure times by marital status, age and sex for the years 1995–1999, extracted from the Norwegian population registers. Mortality data by marital status for Norway are not available for more recent years. We applied the death rates of the 'currently married' to household positions 'cohabiting' or 'married', death rates of the never married to people with household position 'child', 'living alone' or 'other' and death rates of the divorced to 'lone parents'. Since individuals who live in an institution have higher mortality than those who live in private households, death rates for the institutionalized were assumed to be twice as large as those of the never married (given age and sex).

International migration was specified as net immigration in absolute numbers. The level of net immigration for the years 2002–2006 was taken from data that were available from the population registers, and that for later years from Statistics Norway's population forecast; see below. Its distribution by age, sex and household position was borrowed from the previous household forecast.

3.4.3. Consistency

In the LIPRO formulation, the units of analysis are individuals, and household statuses are viewed as their characteristics. This opens up the possibility of inconsistency between the attributes (Van Imhoff and Keilman, 1991; Van Imhoff, 1992). We defined several consistency relationships for events that are experienced by members of the same household, namely for marriage of non-cohabiting partners and of cohabitees, for dissolution of couples and for events that are experienced by children and their parents. For example, the number of non-cohabiting men who marry during a certain period must be equal to the number of non-cohabiting women who marry during that period. Also, the number of men who live with a cohabitee and who experience the exit of the last child (in other words, a change from position cohabiting, with children, to position cohabiting, no children) must be equal to the number of cohabiting women who experience the same event. LIPRO adjusted the parameter values for these household events in such a way that all relationships were fulfilled. We also constrained total numbers of births, deaths and net immigration to corresponding numbers taken from observed population data for the years 2002–2006, and from the medium variant of Statistics Norway's 2005-based population forecast; see `www.ssb.no/emner/02/03/folkfram/arkiv/`. (Statistics Norway published a new population forecast in May 2008, but we could not include that in the present analyses.) Finally, we constrained the total population living in an institution to 41 000, which is the level that was observed in the years 2003, 2004 and 2005. On the basis of the consistent events for household formation, household dissolution, births, deaths and net immigration, LIPRO used bookkeeping equations to update the population by age, sex and household position from one point in time to the next. For instance, the number of married women who were aged 30–34 years on January 1st, 2007, was calculated as the corresponding number for ages 25–29 years 5 years earlier, plus all newly married women and immigrating married women of that age group during 2002–2006, minus all women of that age group who experienced divorce, separation or death during 2002–2006.

Apart from adjustments for consistency, parameters for household events were assumed constant over the forecast period, because the possible trends that were visible in our data were erratic. Thus, the point forecast of shares is equivalent to applying multistate life table transition probabilities to the jump-off population.

3.4.4. *Point forecasts for the shares*

We computed a deterministic household forecast for men and women in 5-year age groups in nine household positions for the years 2007, 2012, ..., 2032 (January 1st) and derived point forecasts $\hat{p}(j, x, s, t)$. In terms of total population, the predicted number is 5.47 million in 2032. This result agrees quite well with the corresponding number that was predicted by Statistics Norway in its 2005-based population forecast, which is 5.43 million. The two numbers are close, as expected, because numbers of births, deaths and net immigrations for the period 2002–2032 were constrained to be equal between the two forecasts; see Section 3.4.3. The remaining small difference is due to different jump-off populations. Since consistency for numbers of births, deaths and net immigrations was not specific for age groups, the age structure that we predict is different from that of Statistics Norway. For example, for the year 2032 we find predicted shares in age groups 0–19, 20–64 and 65 years and older equal to 23%, 56% and 22% respectively. Statistics Norway predicted 28%, 52% and 20% respectively.

Fig. 1 plots the marginal point forecasts for the shares of eight private household positions, irrespective of age and sex. The year 1990 (census information) has been added for reference. Fig. 1 shows a continuation of changes in household and family structures that have gone on for several decades. It becomes increasingly more likely for Norwegian men and women to live alone or in consensual union, and less likely to live with a spouse and children. Data before 1990 that are comparable with those plotted in Fig. 1 are not as detailed as current data, but some comparisons can be made. For instance, whereas 7% of the population lived alone in 1970, the proportion had increased to 10% in 1980 and 14% in 1990 (Statistics Norway (2000), Table 1.3). Our forecast gives 20% in 2032. In 1980, 31% of Norwegian households consisted of a married couple with one or more children; the share was 23% in 1990, and a mere 16% in 2001. The

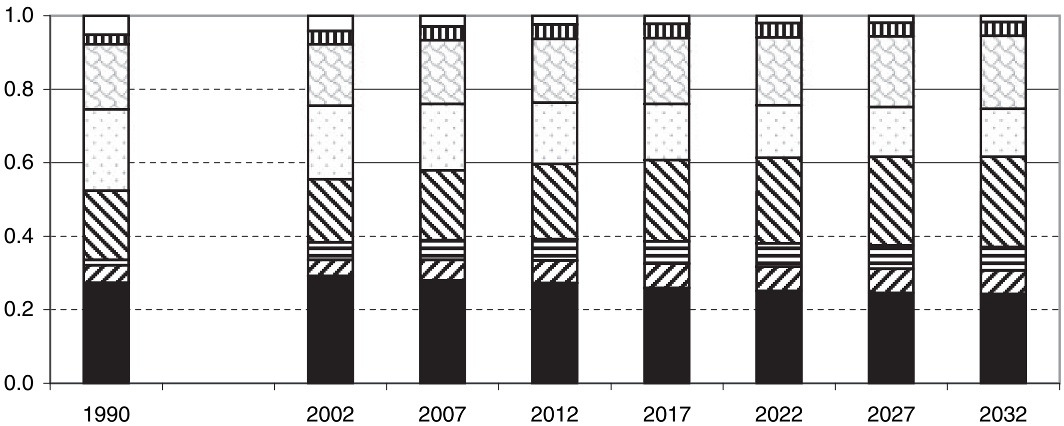

Fig. 1. Shares for eight household positions: ■, child; ▤, cohabiting, with children; ⚄, married, with children; ▥, lone parent; ▨, cohabiting, no children; ▧, married, no children; ⬚, living alone; □, other

result for 2032 in Fig. 1 is 6%. In the three years 1980, 1990 and 2001, cohabiting couples with children made up 1%, 3% and 5% of all households respectively; see www.ssb.no/english/subjects/02/01/20/familie_en/tab-2008-04-03-03-en.html. Our point prediction for 2032 is 6%. In 1977, only 12% of women who were aged 20–24 years lived in cohabitation. Nowadays, the proportion is roughly a third as measured by interview-based survey data (www.ssb.no/english/subjects/02/01/20/samboer_en/tab-2008-04-17-01-en.html), and slightly over a sixth according to census data. (The difference is explained by different definitions of an individual's place of residence across data sources; see the discussion in Section 2.) Our forecast—with a jump-off population based on census data —predicts a value of 23% in 2032 for this share.

Point forecasts for the shares of women who live alone or with their marital spouse are plotted in Figs 2 and 3. Note that elderly women in Norway are less likely to live alone in the future and more likely to live with their husbands. The same trend applies to shares of elderly men. To a large extent, this is due to improved longevity, but also to steeper increases in life expectancy for men than for women (Meslé, 2004). A second important factor is the behaviour of cohorts with people who were born in the 1930s and 1940s. These people were more likely to marry than those in older cohorts, who suffered from the economic crisis in the 1930s and World War II (Keilman and Christiansen, 2008).

3.5. Empirical errors in the shares

To estimate the level of uncertainty that we might expect in a new forecast of shares, we computed empirical errors in the shares $\hat{\rho}(j,x,s,t)$ from the deterministic forecast of Keilman and Brunborg (1995), with jump-off time December 31st, 1990. We analysed six predicted shares for child (CHLD), cohabiting (COH), living with spouse (MAR), living alone (SIN0), living as a single parent together with one or more children (SIN+) and other (OTH). The abbreviations denote individual household positions and will be used later. The shares were evaluated against estimated shares from the 1997 wave and the 2002 wave of the SLC, and data from the census on November 3rd, 2001. As opposed to the deterministic household forecast in Section 3.4, we have not distinguished childless married or cohabiting couples from those who live with children, because the census tables do not give that information. The six shares add up to 1. Hence we

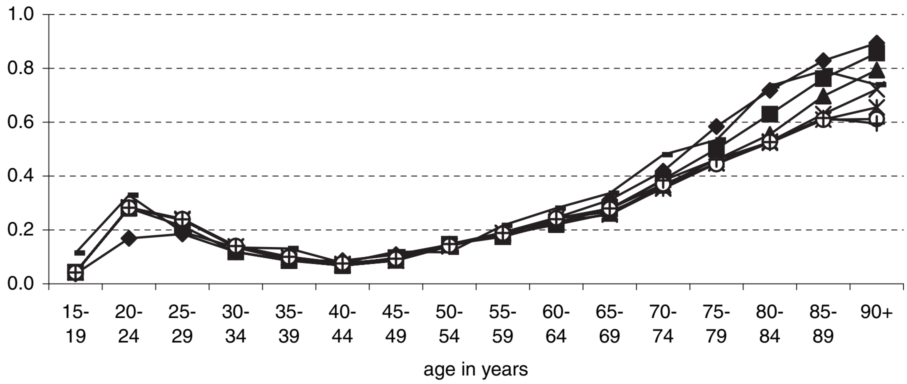

Fig. 2. Proportion of women living alone: ▬, 1990; ◆, 2002; ■, 2007; ▲, 2012; ×, 2017; ⚹, 2022; ○, 2027; +, 2032

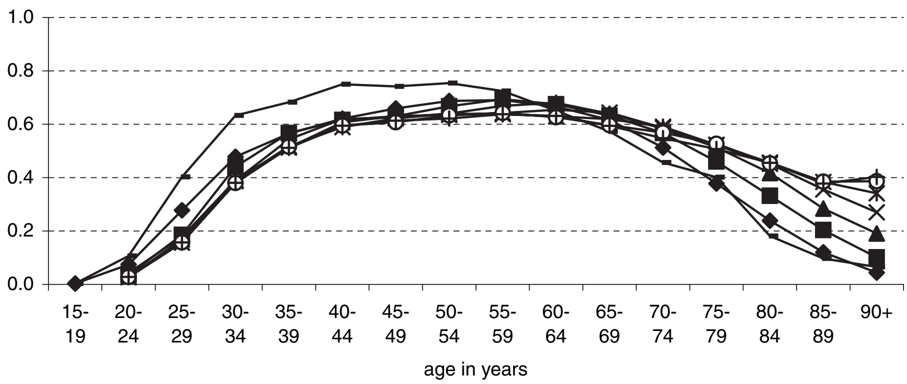

Fig. 3. Proportion of women living with their spouse: ▬, 1990; ◆, 2002; ■, 2007; ▲, 2012; ×, 2017; ⚹, 2022; ○, 2027; +, 2032

excluded people who were in institutional households, for reasons that were mentioned in Section 2.2. Possible consequences of this decision are briefly discussed in Section 4.

Figs 4–6 compare, for cohabiting, married and living alone, predicted shares with the observed shares. The forecast predicted too few young cohabitants (Fig. 4) and young adults living alone (Fig. 6), as opposed to too many children living with their parents. Clearly, the age pattern of the forecast tracks better the census figures than those of the SLC. We suspect that the difference is related to the way that respondents have interpreted their household position in the survey as compared with the census. There is evidence that errors increase over time.

The group ‘other’ consists primarily of adults in ages 25–30 years who live with their parents, or of unrelated people who share a dwelling. Thus, the dividing line between statuses child and other is unclear.

3.6. Modelling errors in the forecasts of shares

We computed errors in share forecasts for three dates (SLC waves of spring 1997 and spring 2002, and the census of November 2001), two sexes, six household positions and 16 5-year age groups: 15–19, 20–24, . . . , 90 years and over. This makes a total of 576 possible errors. In practice, we analysed 385 observations of non-zero shares for the three time points, by age and sex: 12 for child, 81 both for cohabiting and for married, 88 for living alone, 49 for lone parent

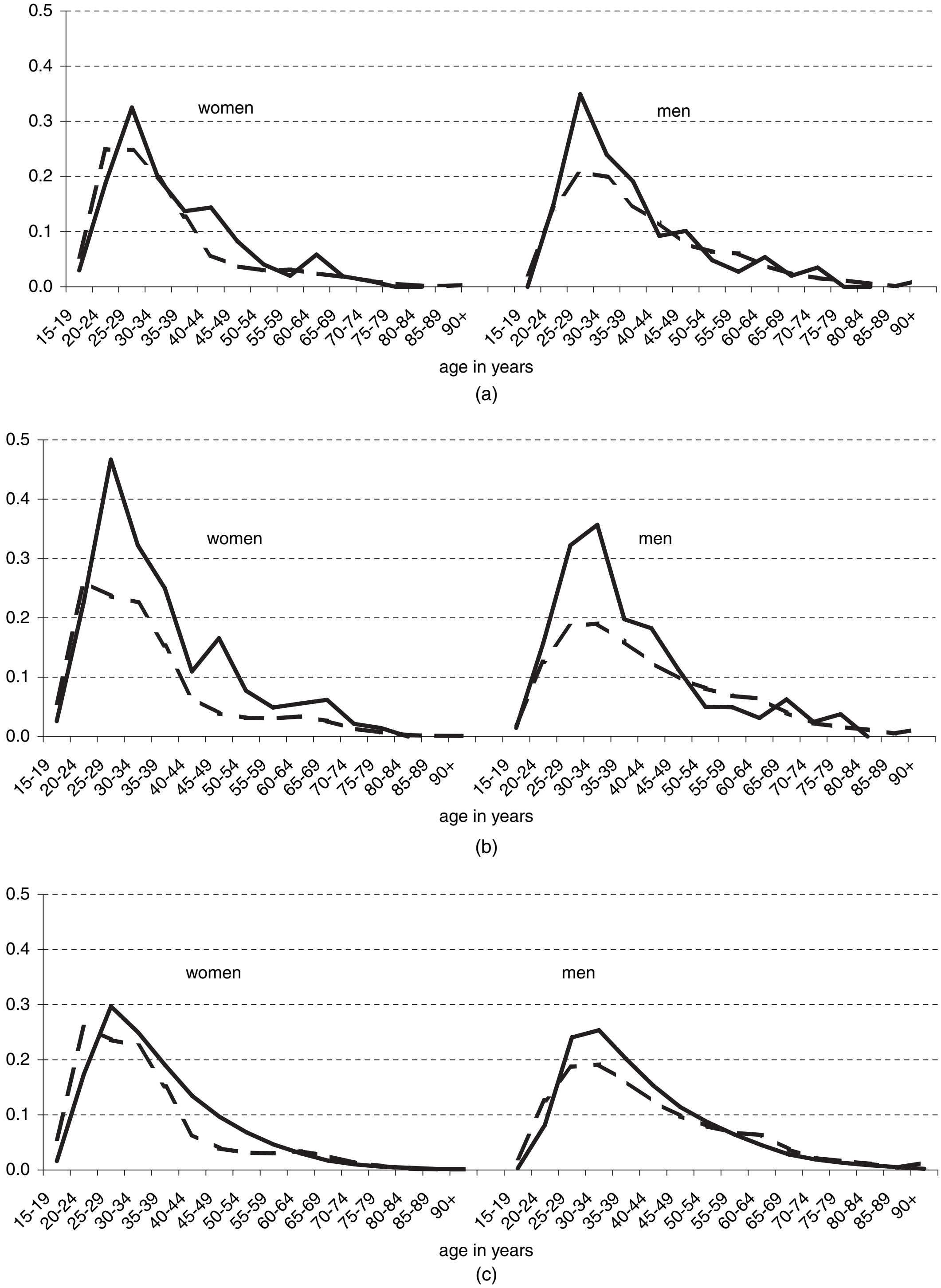

Fig. 4. Observed (———) and predicted (— —) shares for household position 'cohabiting': (a) spring 1997; (b) spring 2002; (c) November 2001

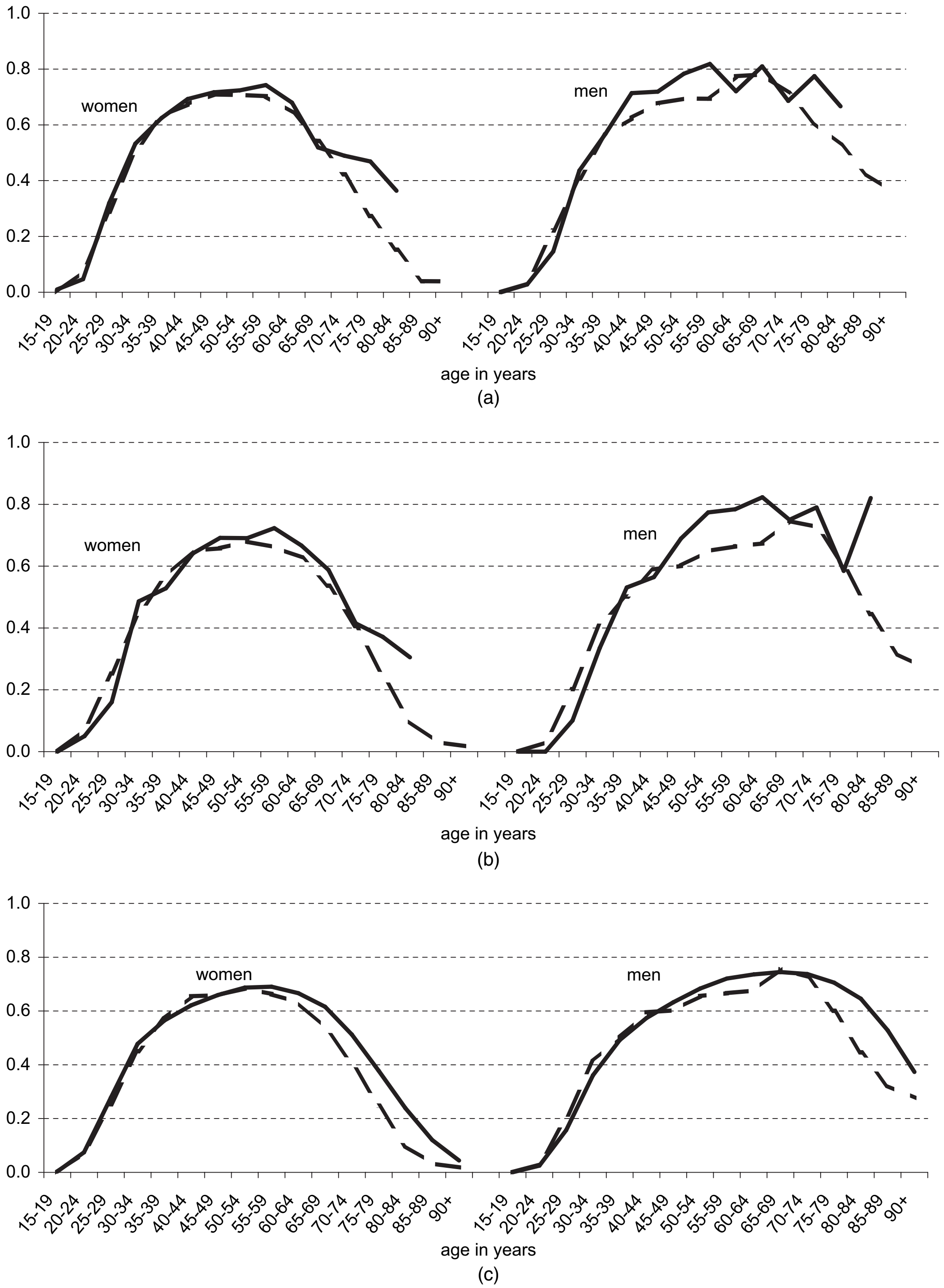

Fig. 5. Observed (———) and predicted (— —) shares for household position 'living with spouse': (a) spring 1997; (b) spring 2002; (c) November 2001

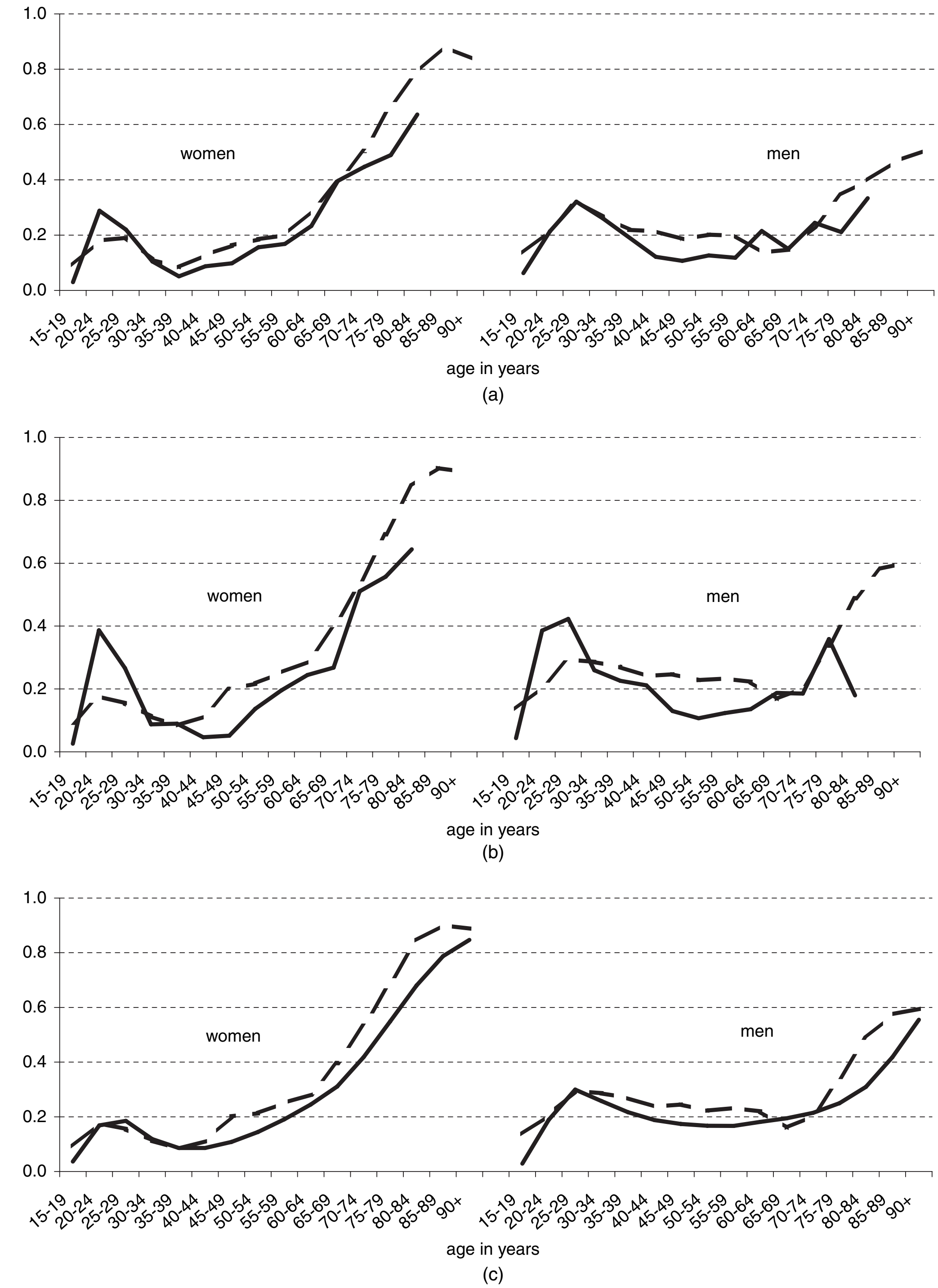

Fig. 6. Observed (———) and predicted (— —) shares for household position 'living alone': (a) spring 1997; (b) spring 2002; (c) November 2001

Table 1. MSEs in predicted shares for five household positions (multinomially logit transformed, with household position living alone as reference category)†

	CHLD	*COH*	*MAR*	*OTHR*	*SIN+*
SLC spring 1997	0.779	0.231	0.199	0.601	2.422
Census November 2001	1.324	0.379	0.278	0.725	1.620
SLC spring 2002	1.868	0.630	0.537	1.488	1.776

†Abbreviations for household positions: CHLD, child living with parents; COH, living in a consensual union; MAR, married, living with spouse; OTHR, living in other private household; SIN+, lone parent.

and 74 for other. We disregarded people who were recorded as child who were age 25 years or older and lone parents age 60 years or more. It quickly turned out that the error patterns in SLC data were highly erratic. However, the survey data were useful in helping us to check that a random-walk model is an acceptable approximation for the increase over time in forecast error (a perfect random walk would have implied a linear increase with time in the variance of the forecast error; for example see Alho and Spencer (2005)). For instance, in a preliminary stage we analysed a multinomial logit transformation for the six shares with the position living alone as the reference category. Write the six shares as $p_1, p_2, \ldots, p_6$, with p_1 the share for living alone, where we suppress indices for time, age and sex. The sum of all shares equals 1. The multinomial logit transformation is $q_i = \ln(p_i/p_1)$, $i = 2, 3, \ldots, 6$. We computed errors in predicted values of q_i according to the old household forecast by comparing predicted q_i-values with observed q_i-values from the census and the two SLC waves. For each of the five household positions and each of the three time points we computed the mean-squared error (MSE) in q_i across age and sex. Table 1 gives MSE values, which can be interpreted as conservative estimates of error variances (since the MSE of a forecast equals the variance of the forecast error plus forecast bias squared). In computing the values in Table 1, we have subtracted the sampling error in 1997 and 2002 from the total forecast error (details are available from the second author on request).

For all household positions except lone parent, the MSE values in Table 1 increase between the three dates, albeit somewhat irregularly. For instance, during the period from the spring of 1997 to November 2001 (approximately 4 years), the MSEs of household positions COH, MAR and OTHR increase much more slowly than they do during the period from November 2001 to the spring of 2002 (about 6 months). Because of the irregularity in the SLC errors that was noted earlier, the estimation of the variances of forecast error was based on the census data alone.

For modelling random evolution of the shares, a logit transformation was first applied. On the basis of Figs 4–6, we opted for a tree-like structure, or a variant of *continuing fractions*, that led to four types of fraction to be modelled (all specific for age, sex and time):

(a) the total share of MAR and SIN0;
(b) the relative share of MAR out of MAR and SIN0;
(c) the relative share of COH out of total share of COH, SIN+, CHLD and OTHR;
(d) the relative share of CHLD and OTHR out of the total share of CHLD, OTHR and SIN+.

Temporarily suppressing indices for age and sex, denote the combined household positions of MAR and SIN0 (type (a) above) as $j = 1$, define the logit transform of the share $\rho(1, t)$ as

$$\xi(1,t) = \ln\left\{\frac{\rho(1,t)}{1-\rho(1,t)}\right\}$$

and write the observed ξ as the sum of the predicted value and an error term: $\xi(1,t) = \hat{\xi}(1,t) + e_{\xi}(1,t)$. Here $\hat{\xi}(1,t)$ denotes the logit transform of the point prediction $\hat{\rho}(1,t)$ in the old household forecast. Next, construct a univariate random walk for $e_{\xi}(1,t)$. In other words, assume the model

$$e_{\xi}(1,t) = \varepsilon(1,1) + \varepsilon(1,2) + \ldots + \varepsilon(1,t),$$

where $\varepsilon(1,t)$ is an independent and identically distributed sequence of random terms with zero expectation and time constant variance to be specified. Next, take the conditional share of MAR out of MAR and SIN0 (type (b) above), compute the logit-transformed value of that share and construct a second univariate random walk in the logit scale. Continue this way with the empirical errors for the remaining two conditional shares. By construction, the four univariate random walks in the logit scale are statistically independent.

We used observed and predicted values of the conditional shares, computed empirical forecast errors for each share and estimated the variances $\mathrm{var}\{\varepsilon(j,t)\}$ as the MSE in the forecast for each of the four shares, assuming that those variances are independent of sex and age. For the four shares that are listed under (a)–(d) above, the estimates turned out to be 0.090, 0.098, 0.129 and 0.321 respectively. We estimated the correlation between the sexes $\mathrm{corr}\{e_{\xi}(j,x,\mathrm{M},t), e_{\xi}(j,x,\mathrm{F},t)\}$ as 0.68, assuming independence of age and household position. The correlation between neighbouring 5-year age groups was 0.65 based on an auto-regressive AR(1) model across age groups, assuming independence across sex and household position. In other words, writing $\xi(j,x,s,t)$ as $\xi(x)$ for simplicity, the AR(1) model that we assumed implies that $\xi(x) = \beta\,\xi(x-1) + u(x)$, $|\beta| < 1$,

Table 2. Average value, coefficients of variation and lower and upper bounds of 80% prediction intervals, for the number of private households, by household type

	Married couple	*One-person household*	*Cohabiting couple*	*Lone parent household*	*Other private household*	*All private households*
2002						
Average	834000	749000	204000	163000	75000	2026000
2010						
Average	877000	828000	267000	184000	42000	2198000
CV (%)	4.2	7.8	10.8	17.2	21.7	1.9
80% low	831000	745000	231000	144000	31000	2146000
80% high	923000	912000	305000	225000	54000	2251000
2020						
Average	919000	926000	318000	200000	37000	2400000
CV (%)	6.8	11.7	15.0	24.6	34.8	3.5
80% low	840000	791000	258000	141000	22000	2293000
80% high	1000000	1068000	380000	264000	54000	2509000
2030						
Average	956000	1044000	345000	207000	36000	2587000
CV (%)	9.3	14.2	18.1	29.7	45.1	5.2
80% low	842000	863000	266000	134000	18000	2418000
80% high	1069000	1242000	428000	288000	58000	2759000

Table 3. Coefficient of variation for predictive distributions of men and women by age and household position, in 2010, 2020 and 2030†

	CVs for the following age groups:		
	20–24 years	*50–54 years*	*80–84 years*
Men, 2010			
MAR	0.312	0.072	0.069
SIN0	0.167	0.217	0.223
COH	0.321	0.288	0.292
CHLD or OTHR	0.106	0.711	0.323
SIN+	1.139	0.331	1.252
Men, 2020			
MAR	0.489	0.113	0.128
SIN0	0.246	0.312	0.337
COH	0.453	0.415	0.389
CHLD or OTHR	0.163	1.585	0.626
SIN+	2.093	0.429	2.176
Men, 2030			
MAR	0.611	0.160	0.205
SIN0	0.311	0.380	0.395
COH	0.559	0.494	0.489
CHLD or OTHR	0.218	1.969	0.878
SIN+	2.822	0.540	2.368
Women, 2010			
MAR	0.300	0.071	0.166
SIN0	0.178	0.230	0.110
COH	0.256	0.280	0.341
CHLD or OTHR	0.150	1.051	0.283
SIN+	0.969	0.258	1.213
Women, 2020			
MAR	0.466	0.113	0.224
SIN0	0.261	0.347	0.194
COH	0.358	0.409	0.440
CHLD or OTHR	0.223	1.886	0.565
SIN+	2.000	0.374	1.857
Women, 2030			
MAR	0.602	0.152	0.280
SIN0	0.329	0.417	0.245
COH	0.444	0.480	0.507
CHLD or OTHR	0.290	2.465	0.962
SIN+	2.252	0.479	1.748

†Abbreviations for household positions: MAR, married, living with spouse; SIN0, living alone; COH, living in consensual union; CHLD or OTHR, child under 25 years of age living with parents, or adult living in other private household; SIN+, lone parent.

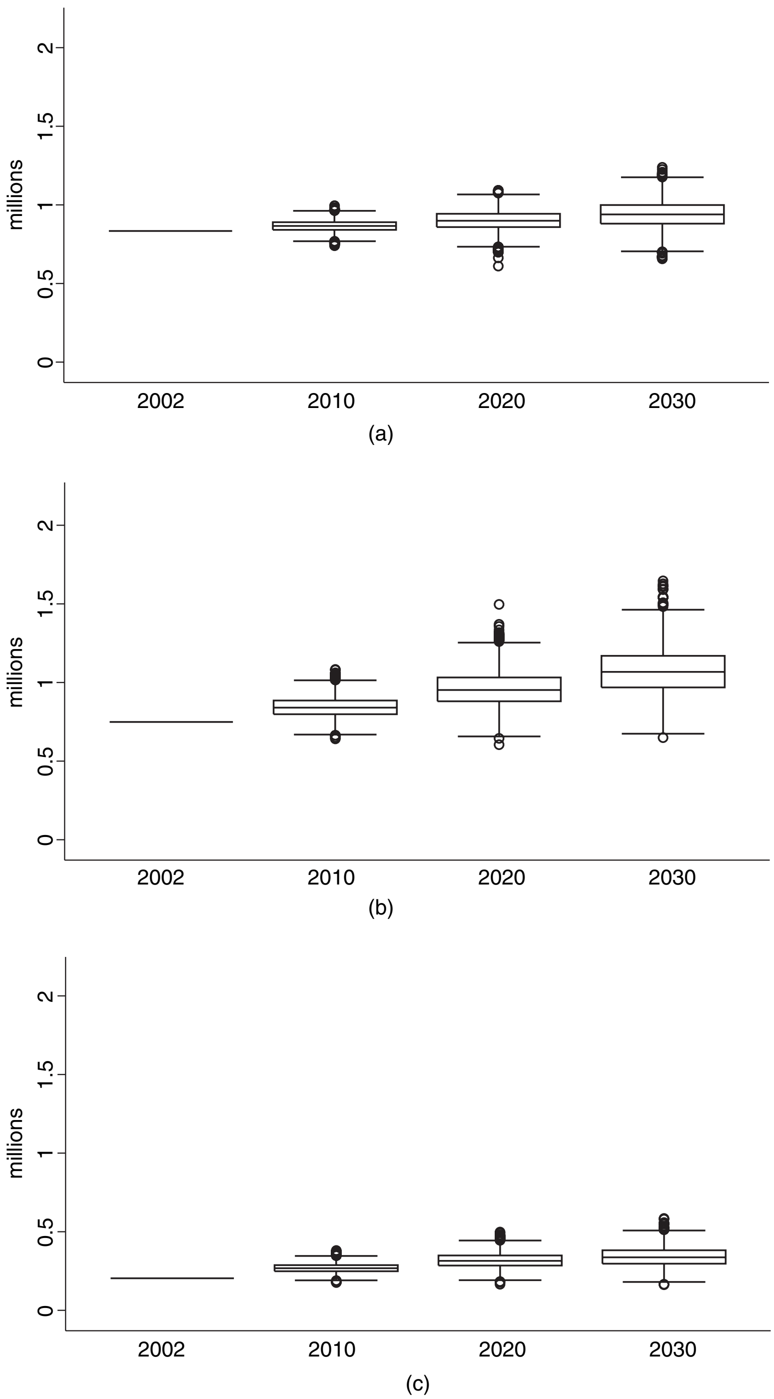

Fig. 7. Box-and-whisker plots of predictive distributions of private households for 2002 (observed), 2010, 2020 and 2030: (a) number of married couple households; (b) number of one-person households; (c) number of cohabiting couple households; (d) number of lone parent households; (e) number of all private households

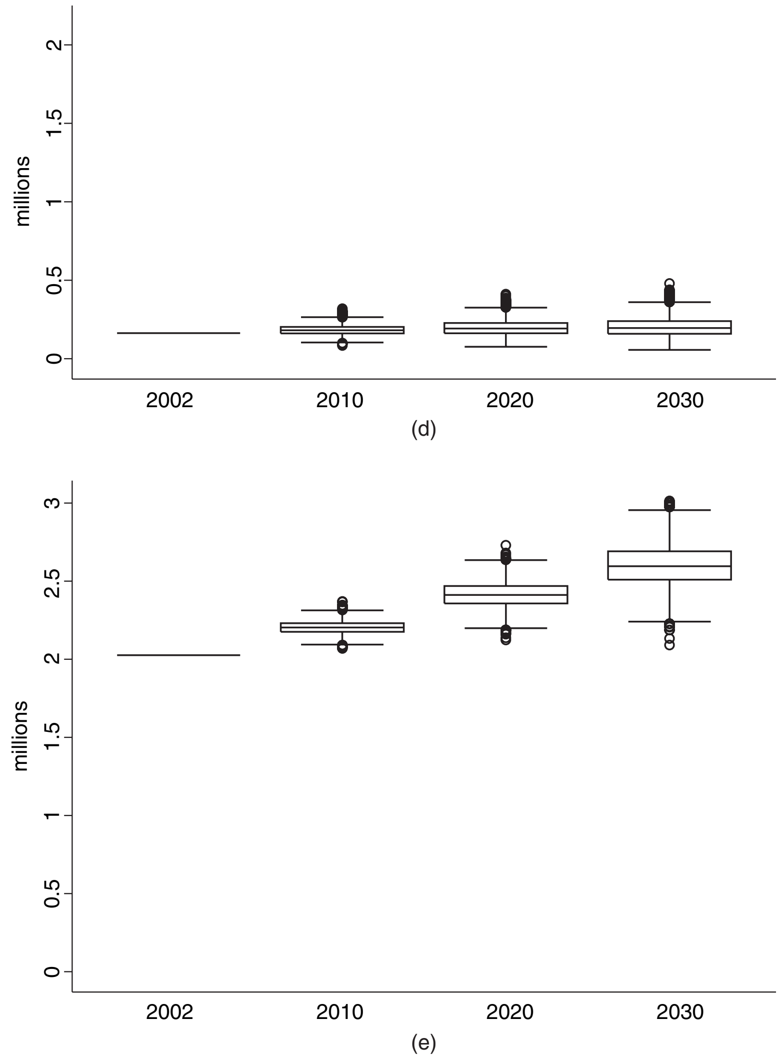

Fig. 7 (*continued*)

where the coefficient β equals the correlation between $\xi(x)$ and $\xi(x-1)$, and the residual $u(x)$ has expectation 0 and variance that is independent of age x.

4. Selected results

We simulated 3000 sample paths of the four random walks that were defined in the previous section, assuming a normal distribution of the $\varepsilon(j,t)$ terms with zero expectations and variances as estimated, added logit-transformed point predictions $\hat{\xi}(1,x,s,t)=\ln[\hat{\rho}(1,x,s,t)/\{1-\hat{\rho}(1,x,s,t)\}]$ from the updated household forecast, transformed back to the ρ-scale and applied the simulated shares to the simulated population numbers $W(x,s,t)$. The four shares cover all private household positions, except for lone parent, whose share is found as 1 minus the sum of the four shares, for each combination of age, sex and time. The shares relate to private households only. In contrast, the population numbers apply to all individuals, both those in private and those in institutional households. This slightly exaggerates our household numbers. For the highest age group (age 90 years and older) our results are not reliable, since 42% of the women and 27% of the men of that age live in an institution, according to the census of 2001. At ages 80–84 years the shares are 10% and 6% respectively. Below age 70 years the shares are below 1%.

For illustration, we computed predictive distributions for the following variables:

(a) the population $V(j,x,s,t)$ for men and women in age groups 20–24, 50–54 and 80–84 years, in 2010, 2020 and 2030, in positions married, living alone, cohabiting, child or other and lone parent;
(b) the number of private households at 2010, 2020 and 2030, for five household types (the numbers of cohabiting and married households were estimated as half the total number of people with that household position; the numbers of one-person and lone parent households equal the numbers of people with that household position; the number of other households was estimated as the number of people with that position (i.e. the number of people with household position child or other who are 25 years or older) divided by 2.5, which was the mean size in the 2002 wave of the SLC; we also computed the total number of private households as the sum of the five households);
(c) the mean size of private households for 2010, 2020 and 2030, computed as the total population $\Sigma_{j,x,s} V(j,x,s,t)$ divided by the number of private households.

In Tables 2 and 3 where relevant results are presented, we shall characterize forecast uncertainty by means of the coefficient of variation CV and the high and low bounds of the 80% prediction intervals for the variables concerned. Equivalently, we could have selected the bounds of 1-standard-deviation or 2-standard-deviation intervals, corresponding to about 68% and 95% coverage probability respectively. However, a coverage probability of 80% has become customary in probabilistic demographic forecasts; see for example Alho and Spencer (2005), chapter 11, Alho *et al.* (2006) and Scherbov and Ediev (2007). The results that we shall present in the form of box-and-whisker plots have the usual first and third quartiles as the upper and lower borders of the boxes. The lines ('whiskers') show the largest or smallest predicted value that falls within a distance of 1.5 times the size of the box from the nearest border. The circles fall farther away—these are considered 'extreme' values.

The number of private households is forecasted to grow by 28% in 2002–2030, from 2.026 million to 2.587 million (Table 2). The population is expected to increase by 16% (not shown here), implying a drop in the mean size of private households from 2.21 to 2.04. The fraction of one-person households in 2030 is expected to be 40%, up from 36% in 2002. Married couple households and one-person households have the lowest relative forecast uncertainty as measured by the coefficient of variation. At the same time, these household types are the most numerous. The lower bound of the 80% interval for the total number of households increases regularly, which indicates that the chance for a systematic decrease in that number is small. In fact, none of the 3000 simulations in any of the three years 2010, 2020 or 2030 show a number that is smaller than that in 2002 (see also Fig. 7(e), which is discussed below). At the same time, the simulations imply an estimated chance of 12% that there will be fewer married couple households in 2030 than there were in 2002 (the chances are 20% in 2010 and 15% in 2020).

Fig. 7 illustrates these predictive distributions graphically in the form of box-and-whisker plots based on the 3000 simulated values. The vertical scales for the five parts are the same, so that uncertainty can be compared between household types. Whereas Table 2 shows modest *relative* uncertainty for one-person households (as measured by CV), the fact that these households are so numerous implies that *absolute uncertainty* is largest for this household type; compare the heights of the boxes.

Table 3 shows the coefficient of variation for the number of men and women in various household positions, for selected age groups. The level of uncertainty is directly related to the variance of the $\varepsilon(j,t)$ terms of the random walk, but this effect is mediated by the point prediction. For instance, married men and women had low values for these variances, but in the 20–24 years age group, where their shares are low, CV is high. In this age group, household position CHLD or OTHR (in practice child) has the lowest uncertainty among the five household positions. In the 80–84 years age group men are likely to live with their spouses (household position MAR), whereas women frequently live alone (SIN0), which results in low CVs. These differences are illustrated by the box-and-whisker plots for shares in Fig. 8.

The results in Table 3 are based on the predictive distributions for numbers of people in various household positions. It is instructive to analyse whether the uncertainty in those predictions derives mainly from that in the predictions for the shares ρ, or from the predictions for the population numbers W. Write the number of people in a given household position, for a certain combination of age, sex and time, as $V=\rho W$ for short. Its coefficient of variation is

$$\mathrm{CV}(V)=\mathrm{var}(V)^{1/2}/E[V].$$

Since ρ and W are assumed independent (see Section 5 for a discussion), $E[V]=E[\rho]\,E[W]$. Using the formula for conditional variances, $\mathrm{var}(V)=E[\mathrm{var}(V|\rho)]+\mathrm{var}(E[V|\rho])=E[\rho^2]\mathrm{var}(W)+E[W]^2\mathrm{var}(\rho)$. This leads, after some manipulation, to the formula

$$\mathrm{CV}^2(V)=\mathrm{CV}^2(\rho)+\mathrm{CV}^2(W)+\mathrm{CV}^2(\rho)\,\mathrm{CV}^2(W).$$

Consider, for example married women aged 50–54 years, in 2030. Estimated values of $\mathrm{CV}^2(\rho)$ and $\mathrm{CV}^2(W)$ are 0.0213 and 0.0018 respectively (not shown here). Plugging in these values in the right-hand side of the equality, we obtain $0.0213+0.0018+0.0213\times 0.0018=0.0231$ for $\mathrm{CV}^2(V)$; compare the $\mathrm{CV}(V)$ value of 0.152 in Table 3. Hence the uncertainty in V, as expressed by the coefficient of variation, derives almost entirely from that of the share ρ. This holds fairly generally for those in the age groups that were already born at the jump-off time. However, in the oldest ages the uncertainty of forecasting mortality is an important factor of uncertainty. For instance, men who are aged 80–84 years in 2030 have $\mathrm{CV}=0.1456$, whereas CV for the married share is 0.1424. Thus, both sources contribute almost evenly to the coefficient of variation of the number of married men that is given as 0.205 in Table 3.

5. Summary and concluding remarks

We have developed a method for computing probabilistic household forecasts and applied it to data from Norway. The method combines a probabilistic population forecast with a probabilistic forecast for the shares of people in six household positions, by age and sex. Conceptual difficulties in defining household positions appear to have an effect on the quality of the basic data. Thus, empirical estimates based on Norwegian survey data differed markedly from the census estimates that were obtained for essentially the same point in time. In this work we primarily relied on the Norwegian census, because it appeared to be more consistent with the earlier forecast that we used in error estimation, and our work was primarily motivated by the economic implications of household structure. However, it is clear that for other applications of household data changing perceptions of one's position may be of equal interest. Such forecasts would be expected to be more uncertain.

Point predictions for the shares were obtained from the multistate LIPRO model. The transition probabilities estimated around year 2000 imply a continued increase in the number of single-person households.

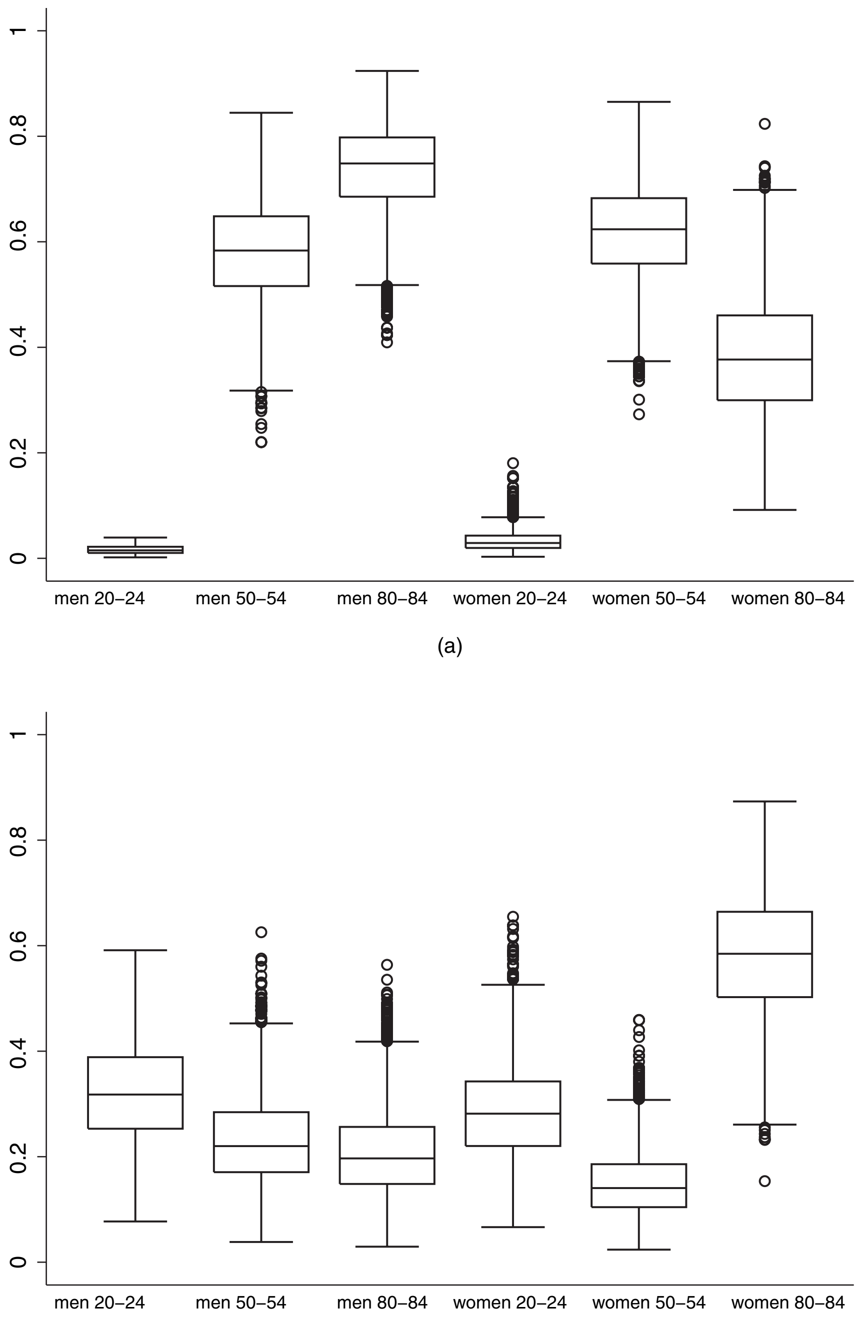

Fig. 8. Box-and-whisker plots for predictive distributions of shares (a) living with spouse and (b) living alone in 2030: men and women in three age groups

Uncertainty parameters were estimated from observed errors of an old household forecast against a subsequent census. Our preferred model for the shares relied on an empirically defined hierarchical model of continued fractions. We also experimented with the direct modelling of the shares but found (the details are not shown) that the resulting predictive distributions for the shares had much larger error variances. It is important to exercise care in the stochastic modelling of the shares.

We find that the coefficient of variation in future numbers of married couples, cohabiting couples and one-person households is lower than that for lone parents and other private households. Moreover, for most ages and most household positions, it is the uncertainty of the share, rather than the uncertainty of the base population, that determines the uncertainty of the number of individuals in different household positions.

There are various sources of uncertainty that have an influence on our results. The deterministic population model and the deterministic household model are both of the bookkeeping type. Hence they will result in perfectly accurate forecasts, once the parameter values have been selected correctly. For LIPRO this is true in so far as we mean by parameter values those that follow implicitly from the consistency adjustments. Observed errors in the household shares come from wrong assumptions in the old deterministic household forecast and errors in the census data. It is reasonable to assume that the assumptions of the old household forecast have a bigger effect than the errors in the census data (although the latter source of error may be considerable for shares of children aged 20–24 years who live in the parental household). The variation in the results of the probabilistic household forecast comes from several sources:

(a) estimated (co)variances in the parameters for fertility, mortality and migration in the population model;
(b) assumptions about the way that these (co)variances propagate over time through the age structure;
(c) estimated (co)variances in household shares;
(d) the assumed random-walk model for these shares.

Our approach excludes other sources of uncertainty. In particular, there is no covariance between population variables and household variables, owing to our implicit assumption of independence. This point can be clarified as follows.

We have simulated random shares that allocate individuals over various household positions. The shares were combined with random population numbers from an existing stochastic population forecast. This approach implicitly assumes independence between the population numbers (and the underlying processes of birth, death and migration) and the shares (and the underlying processes of household formation and dissolution). In reality, the two types of variable are not necessarily independent. For instance, mortality is most probably correlated with the chance to live alone—in particular for the elderly, through widowhood. Other things being equal, when there are many elderly men (for instance because of low mortality), the share of elderly women who live alone is probably low, and vice versa. However, correlations of this and similar kinds are likely to have only limited numerical importance. Intuitively, this is because, first, the uncertainty in population sizes for the elderly, and hence in survival, is much less than that of the shares in most ages. Second, survival of males and survival of females have a high positive correlation (greater than 0.80), so the difference in survival has an even smaller uncertainty than the populations themselves.

A somewhat subtle problem in our stochastic formulation relates to the consistency constraints that are used in LIPRO. These are implemented by an iterative calculation, and it is not practicable to try to enforce them for all sample paths, that numbered 3000 in our application. We investigated the possibility that increasing correlation between shares for males and females might reduce any discrepancy. The results in Appendix A confirm that this is so.

A basic difficulty in the modelling of the household position data derives from the fact that households are viewed as characteristics of the individuals. This approach has been adopted on grounds of conceptual simplicity. Individuals enter and leave the household, and this changes its structure and type. However, since the behaviour of two individuals may be linked, we are confronted with the so-called two-sex problem, for instance in the case of couple formation. A challenging future direction is the development of a richer population model that includes actual households as units of observation.

The data requirements for our approach are not very strong. Therefore, the method that we report here can also be applied to other countries. First, a probabilistic forecast of the population by sex and age is required. The approaches that were developed in the past to compute such forecasts have been applied to numerous countries—most of these in the western world (cf. the examples that were mentioned in Section 1), but not exclusively. See Booth (2006) for a review of methods. Second, point predictions and uncertainty measures are required for the random shares that distribute the population over several household positions. The point predictions can be derived from any household forecast model that predicts the population broken down by age, sex and household position. We have used LIPRO, which requires data on household events, among others. Other models which demand less detailed data have been employed elsewhere (e.g. ProFamy (see Zeng *et al.* (2007)), or the method that has been developed by Statistics Netherlands (see Alders and Manting (1998)). We estimated uncertainty parameters for the shares from the errors in an old household forecast. Those errors require observed data on household shares by age and sex, and an old forecast for those shares. Population data with information on age, sex and household position from the 2000 round of population censuses in European Union member countries are available from Eurostat. Similar data should exist for other countries as well, because the United Nations Economic Commission for Europe recommended that this type of information should be included in the set of tables which each member country should produce in the 2000 round of population censuses (United Nations Economic Commission for Europe, 1998). Thus it should be possible to evaluate household forecasts that were made before 2000 for the countries where such forecasts have been computed. If they do not exist, error estimates can be borrowed from other countries. Of course, whenever time series of past estimates of shares are available, error estimates can be derived directly from the statistical modelling of such series.

The method in which random shares are combined with random numbers for the population is not restricted to the particular example in this paper. A similar method could be applied to probabilistic forecasts for other divisions of the population, e.g. household size, health or disability status, region of residence and labour market status. For instance, Scherbov and Ediev (2007) applied random headship rates (the share of the population that is a head of a household) to a probabilistic population forecast. Their headship rates are specific for household size, and thus they could predict households by size. Also, statistical agencies of many countries compute regional forecasts regularly. The errors in historical regional forecasts can be analysed, modelled and predicted. Combined with regional shares in the current regional forecast those errors give random regional shares, which, in turn, can be applied to a probabilistic population forecast.

Acknowledgements

Part of this project was carried out when the authors participated in the international research

Table 4. Average and standard deviation of the difference between the number of married and cohabiting men and women

	Results for correlation 0.68		*Results for correlation 0.75*		*Results for correlation 0.95*	
	Married	*Cohabiting*	*Married*	*Cohabiting*	*Married*	*Cohabiting*
2010						
Average	−11023	1443	−10992	1439	−10883	1433
Standard deviation	641	490	581	439	357	235
2020						
Average	−19000	−3140	−18933	−3241	−18672	−3623
Standard deviation	1010	778	915	696	564	371
2030						
Average	−31902	3834	−31775	3654	−31302	2967
Standard deviation	1388	1017	1262	910	802	484

group ‘Changing family patterns in Norway and other industrialized countries’ at the Centre for Advanced Study, the Norwegian Academy of Science and Letters, Oslo, during the academic year 2006–2007. Øystein Kravdal kindly supplied us with data on deaths and exposure times by marital status, age and sex for the years 1995–1999. We also gratefully acknowledge useful comments that were received from two reviewers and the Joint Editor of the journal.

Appendix A

LIPRO enforces consistency restrictions for the point forecast, but this cannot be easily extended to the 3000 simulation rounds. To see the effect of cross-sex correlation on lack of consistency regarding the number of marred and cohabiting couples, we increased the correlation from the reference value of 0.68 to 0.75 and 0.95 (Table 4).

An increase in correlation decreases the average difference marginally, and the standard deviation substantially. Differences between men and women increase with lead time, and the standard deviations indicate that the expected differences are significantly different from 0. Compared with the numbers in Table 2, the differences are small.

References

Alders, M. (1999) Stochastische huishoudensprognose 1998-2050 (Stochastic household forecast 1998-2050). *Maandstatist. Bevolkng*, no. 11, 25–34.

Alders, M. (2001) Huishoudensprognose 2000-2050: veronderstellingen over onzekerheidsmarges (Household forecast 2000-2050: assumptions on uncertainty intervals). *Maandstatist. Bevolkng*, no. 8, 14–17.

Alders, M., Keilman, N. and Cruijsen, H. (2007) Assumptions for long-term stochastic population forecasts in 18 European countries. *Eur. J. Popln*, **23**, 33–69.

Alders, M. and Manting, D. (1998) Household scenarios for the European Union. *Maandstatist. Bevolkng*, no. 10, 11–27.

Alho, J. M. (1998) *A Stochastic Forecast of the Population of Finland: Reviews 1998/4*. Helsinki: Statistics Finland.

Alho, J. M., Alders, M., Cruijsen, H., Keilman, N., Nikander, T. and Pham, D. Q. (2006) New forecast: population decline postponed in Europe. *Statist. J. UN ECE*, **23**, 1–10.

Alho, J., Cruijsen, H. and Keilman, N. (2008) Empirically based specification of forecast uncertainty. In *Uncertain Demographics and Fiscal Sustainability* (eds J. Alho, S. Hougaard Jensen and J. Lassila), pp. 34–54. Cambridge: Cambridge University Press.

Alho, J. and Spencer, B. (2005) *Statistical Demography and Forecasting*. New York: Springer.

Booth, H. (2006) Demographic forecasting: 1980 to 2050 in review. *Int. J. Forecast.*, **22**, 547–581.

Conference of European Statisticians (2006) *Recommendations for the 2010 Censuses of Population and Housing*. Geneva: United Nations.
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*, working paper 45. Geneva: Economic Commisson for Europe.
Glaser, K., Grundy, E. and Lynch, K. (2003) Transitions to supported environments in England and Wales among elderly widowed and divorced women: the changing balance between co-residence with family and institutional care. *J. Wom. Agng*, **15**, 107–126.
Grundy, E. (2001) Living arrangements and the health of older persons in developed countries. *Popln Bull. UN*, **42–43**, 311–329.
Hoffmann-Nowotny, H. J. (1987) The future of the family. In *Plenaries of the European Population Conference*, pp. 113–200. Helsinki: International Union for the Scientific Study of Population–Central Statistical Office of Finland.
Holmberg, I. (1987) Household change and housing needs: a forecasting model. In *Family Demography—Methods and Their Application* (eds J. Bongaarts, T. Burch and K. Wachter), pp. 327–341. Oxford: Clarendon.
Jiang, L. and O'Neill, B. (2004) Toward a new model for probabilistic household forecasts. *Int. Statist. Rev.*, **72**, 51–64.
Jiang, L. and O'Neill, B. (2006) Impacts of demographic events on US household change. *Interim Report IR-06-030*. Laxenburg: International Institute for Applied Systems Analysis.
Keilman, N. and Brunborg, H. (1995) Household projections for Norway, 1990-2020. *Report 95/21*. Oslo: Statistics Norway.
Keilman, N. and Christiansen, S. (2008) Norwegian elderly less likely to live alone in the future. To be published.
Keilman, N. and Keyfitz, N. (1988) Recurrent issues in dynamic household modelling. In *Modelling Household Formation and Dissolution* (eds N. Keilman, A. Kuijsten and A. Vossen), pp. 254–285. Oxford: Clarendon.
Keilman, N., Kuijsten, A. and Vossen, A. (eds) (1988) *Modelling Household Formation and Dissolution*. Oxford: Clarendon.
Keilman, N., Pham, D. Q. and Hetland, A. (2002) Why population forecasts should be probabilistic—illustrated by the case of Norway. *Demogr. Res.*, **6**, 409–454.
King, D. (1999) Official household projections in England: methodology, usage and sensitivity tests. *Joint Economic Commission for Europe–EUROSTAT Work Session on Demographic Projections, Perugia, May 3rd–7th*, working paper 47. Geneva: Economic Commission for Europe.
Lee, R. (1999) Probabilistic approaches to population forecasting. *Popln Devlpmnt Rev.*, **24**, suppl., 156–190.
Lee, R. and Tuljapurkar, S. (1994) Stochastic population forecasts for the United States: beyond High, Medium, and Low. *J. Am. Statist. Ass.*, **89**, 1175–1189.
Lenoir, R. (2007) The family as a social institution: struggles over legitimate representations of reality. In *Symbolic Power in Cultural Contexts: Uncovering Social Reality* (eds J. Houtsonen and A. Antikainen), pp. 31–41. Rotterdam: Sense.
Lutz, W., Sanderson, W. and Scherbov, S. (1996) Probabilistic population projections based on expert opinion. In *The Future Population of the World: What can We Assume Today?* (ed. W. Lutz), pp. 397–428. London: Earthscan.
Lutz, W., Sanderson, W. and Scherbov, S. (2001) The end of world population growth. *Nature*, **412**, 543–545.
Lutz, W. and Scherbov, S. (1998) An expert-based framework for probabilistic national population projections: the example of Austria. *Eur. J. Popln*, **14**, 1–17.
McMillan, D. B. and Herriott, R. (1985) Toward a longitudinal definition of households. *J. Econ. Socl Measmnt*, **13**, 349–360.
Meslé, F. (2004) Espérance de vie: un avantage féminine menacé? *Popln Soc.*, no. 402, 1–4.
Moen, Ph. and Forest, K. B. (1999) Strengthening families: policy issues for the twenty-first century. In *Handbook of Marriage and the Family*, 2nd edn (eds M. B. Sussman, S. K. Steinmetz and G. W. Peterson), pp. 633–663. New York: Plenum.
Muller, C., Gnanasekaran, K. and Knapp, K. (1999) *Housing and Living Arrangements for the Elderly: an International Comparison Study*. New York: International Longevity Center.
National Research Council (2000) *Beyond Six Billion: Forecasting the World's Population* (eds J. Bongaarts and R. Bulatao). Washington DC: National Academy Press.
Normann, T. M. (2004) Samordnet levekårsundersøkelse 2002—panelundersøkelsen (Coordinated survey on living conditions 2002—panel survey). *Dokumentasjonsrapport*. Statistics Norway, Oslo.
O'Neill, B. and Chen, B. (2002) Demographic determinants of household energy use in the United States. In *Popln Devlpmnt Rev.*, **28**, suppl., 53–88.
Scherbov, S. and Ediev, D. (2007) Probabilistic household projections based on an extension of headship rates method with application to the case of Russia. *Joint Economic Commission for Europe–EUROSTAT Work Session on Demographic Projections, Bucharest, Oct. 10th–12th*, working paper 16. Geneva: Economic Commission for Europe.
Schoen, R. (1988) *Modeling Multigroup Populations*. New York: Plenum.
Settles, B. (1999) The future of the families. In *Handbook of Marriage and the Family*, 2nd edn (eds M. B. Sussman, S. K. Steinmetz and G. W. Peterson), pp. 143–175. New York: Plenum.

Smith, R. T. (1968) Family: I, comparative structure. In *International Encyclopedia of the Social Sciences*, vol. 5 (ed. D. Sills), pp. 301–313. New York: McMillan and Free Press.
Statistics Norway (2000) *Social Trends 2000*. Oslo: Statistics Norway.
Törnqvist, L. (1949) Om de synspunkter, som bestämt valet av de primäre prognosantagendena (On the points of view that determined the choice of the main forecast assumptions). In *Beräkninger Rörende Finlands Befolkning, dess Reproduktion och Framtida Utveckling* (eds J. Hyppolä, A. Tunkelo and L. Törnqvist), pp. 69–75. Helsinki: Statistiska Centralbyrån.
United Nations (1973) *Methods of Projecting Households and Families*. New York: United Nations.
United Nations Economic Commission for Europe (1998) *Recommendations for the 2000 Censuses of Population and Housing*. Geneva: United Nations.
US National Resources Planning Committee (1938) *The Problems of a Changing Population*. Washington DC: Government Printing Office.
Van Imhoff, E. (1992) A general characterization of consistency algorithms in multidimensional demographic projection models. *Popln Stud.*, **46**, 159–169.
Van Imhoff, E. and Keilman, N. (1991) *LIPRO 2.0: an Application of a Dynamic Demographic Projection Model to Household Structure in The Netherlands*. Amsterdam: Swets and Zeitlinger.
Zeng, Y., Land, K., Wang, Z. and Gu, D. (2007) U.S. family household momentum and dynamics: an extension and application of the ProFamy method. *Popln Res. Poly Rev.*, **25**, 1–41.