Journal of Official Statistics, Vol. 13, No. 3, 1997, pp. 245–277

Ex-Post Errors in Official Population Forecasts in Industrialized Countries

Nico Keilman[1]

The existing literature on ex-post errors observed for historical population forecasts made by statistical agencies in 16 industrialized countries is reviewed. The amount and type of data in these studies (total population size, age structure, population growth rate, components of change), the level of detail (numbers of births and deaths, crude rates, total fertility rates, life expectancies), and the period covered vary considerably among the countries. Attention is given to forecast accuracy for total population size, population growth rates, fertility, mortality, the age structure, and the dependency ratio. Among the issues covered are relative errors for fertility and mortality, and common patterns across the countries. On the basis of a data set from Norway (covering the period 1969–1989, forecasts made between 1969 and 1987) and one from the Netherlands (for the period 1950–1986, forecasts made between 1950 and 1980) we investigate a possible correlation between forecast errors for fertility and mortality, and a possible reduction in forecast errors over time. The article concludes with suggestions for including results from ex-post evaluations in official population forecasts.

Key words: Demography; forecasting; population projections; accuracy; APC-model.

1. Introduction

Users who work with the results of population forecasts should be aware of the uncertainty in the figures. A school planning agency, e.g., should not work with just one figure for the future number of children in school ages in some future year. They must be prepared for deviations from the forecasted numbers of children and take alternative trajectories into account. Not only should forecast users have an understanding of the fact that population estimates are inherently uncertain, but also of the *degree* of that uncertainty. Ideally, they should be supplied not only with point predictions, but also with confidence intervals around these predictions.

Most official population forecasts published by statistical agencies fail to provide satisfactory information on the uncertainty of the forecast to the users. The uncertainty is reflected in the presentation of several alternatives, typically a high, a medium, and a low forecast variant. From each of these alternatives a user could choose predicted numbers and investigate the respective effects of different prospective population trends on the plans he or she is to draw up.

[1] Statistics Norway, P.O.Box 8131 Dep, N-0033 Oslo 1, Norway.

Acknowledgment: This is the revised version of a paper presented at the Nordic Symposium on Analysis of Errors in Demographic Forecasts with Implications on Policy, Koli, Finland, 30 March–2 April 1995. Helpful comments on the earlier version from Juha Alho, Wolfgang Lutz, Inger Texmon, as well as three journal referees are gratefully acknowledged. Thanks are also due to Hanne Spøhr, who provided me with unpublished data regarding the accuracy of forecasts made by Statistics Denmark since 1970.

The predictions will give him or her the ratio between the high and the low alternatives. But it is not sufficient to know that the future number of pupils according to the high forecast variant will be x per cent higher than with the low variant. The forecast will be of much better use if the probability is known with which the future population figures will fall between the high and the low forecast alternatives. School planning authorities may have to take an entirely different decision if the probability is 90 per cent that the true future population figure falls between the high and the low alternatives than if the probability is 30 per cent. In other words, uncertainty should be quantified.

Theoretically speaking, a stochastic model of the cohort-component type could be used to quantify uncertainty, and to compute confidence intervals around point forecasts. The earliest attempts to take forecast uncertainty into account on the basis of such a model dealt with random fluctuations only (Pollard 1966; Schweder 1971). However, the most important source of error in population forecasts is the *level* of the fertility, migration and mortality parameters, not their randomness. Other models rest on implausible assumptions, such as stochastic stationarity in the fertility and mortality rates (Cohen 1986). Or they only give an indication of the uncertainty around total population size (De Beer 1992; Stoto 1983; Keyfitz 1981), whereas most users are interested in specific age groups. Models which do not have the kind of drawbacks referred to here are relatively complicated (Alho and Spencer 1995; Lee and Tuljapurkar 1994), and much work remains to be done before such models can be used for official forecasts produced by statistical agencies.

Another approach, which is the perspective taken in this article, is to focus on analysing ex-post observed errors in historical forecasts. Given a long enough record of old forecasts, these can be compared with observed population numbers, and the analysis of such numbers mirrors the errors in *previous* forecasts. The results of such analysis can be used to give an indication of the uncertainty connected to *future* forecasts. The drawback of this empirical approach compared to the stochastic model approach is that one lacks a model that can be thought of as having generated the errors. This may make it difficult to find a possible structure and regularity in observed errors. But the advantage of the empirical approach is that forecast errors (albeit old forecasts) can be computed immediately without taking recourse to specific assumptions. It will be clear that the two approaches are complementary. Data gathered by the empirical approach may be used to fit the error model and assess its quality. Next, the model may be used to extrapolate predicted errors from the domain of historical forecasts to that of future ones.

The aim of the current article is to synthesize the literature on ex-post observed errors in official national population forecasts in industrialized countries, and furthermore to report on a few additional analyses carried out on the basis of two, relatively detailed, data sets with ex-post observed forecast errors for Norway and the Netherlands. The literature review is contained in Section 2. We look at errors in total population, population growth rate, population structure by age and sex, and fertility and mortality indicators (number of births, crude birth rate, TFR, number of deaths, crude death rate, life expectancy). No published analyses that evaluate the accuracy of forecasts for the dependency ratio are known, and therefore we include in Section 2 some results for the accuracy of this indicator that we believe are new. These results are based on data sets for Austria, Norway, and the Netherlands. The review does not explicitly take up international migration, because the very few

studies that have analysed forecast errors for this component have not found much regularity in the error patterns. Section 3 investigates a possible correlation between errors in births forecasts and those in deaths forecasts for Norway (forecasts made between 1969 and 1982) and the Netherlands (forecasts made between 1950 and 1980). Such a correlation, which is important for simplifying stochastic models that describe the propagation of errors has not been analysed earlier on the basis of ex-post observed errors. Next we look in Section 4 at a possible improvement of forecast accuracy for these two countries. If more recent forecasts are more accurate than earlier ones, this implies that parameters for forecast error models are not stationary. Finally (Section 5) we formulate a number of recommendations as to how statistical agencies can include results from ex-post evaluations in their forecast reports.

Most statistical agencies prefer to speak of a *projection*, which indicates a purely conditional computation: how would the population structure evolve in the future if certain developments in fertility, mortality, and international migration took place? One may argue that an error analysis is useless in that case, because a projection is always 100 per cent accurate (except for computation errors). However, most users interpret the projection results as a *forecast*, indicating a likely development, given the current knowledge of the forecaster. Therefore an assessment of the accuracy of those future population numbers would be important from the user's point of view. The results of a high and a low variant can also be evaluated, provided these variants can be interpreted as *uncertainty variants*, which indicate a margin around a most likely medium variant. If, on the other hand, the interpretation is that of *genuinely alternative futures*, the variants should be interpreted as projections, and an error analysis is less interesting.

The analyses are restricted to official demographic forecasts and projections as these are routinely being produced by statistical agencies. Two arguments have guided this choice: (i) official demographic forecasts, as opposed to products of individual researchers, are widely used in government planning; (ii) the institutional base of these forecasts guarantees a certain continuity in the production of the forecast, on the basis of which we may analyse *a series of* historical forecasts (in order to discover a possible structure in the forecast errors), and, moreover, which makes it worth while formulating recommendations for handling forecast uncertainty by the statistical agencies.

2. Earlier Findings

Most forecast *users* interested in the accuracy of past population forecasts at the national level will be satisfied with error statistics for the total population, the growth rate and the age structure. However, *forecast producers* are (or should be) interested in the reasons why old forecasts went wrong, and then one has to move from the level of forecast results to that of the assumptions concerning components of growth. The standard method for producing national population forecasts, applied by all industrialized countries, is the cohort-component method (Cruijsen and Keilman 1992, p. 3). The model involved is basically a demographic accounting framework that allows the forecaster to assess the consequences of an assumed set of age-specific fertility, mortality, and external migration patterns. Thus, ideally one should carry out an error analysis for these age-specific indicators, but this information is usually no longer available for forecasts made in the past. Summary

Table 1. Evaluation studies of national population forecasts made by statistical agencies

	Reference	Forecast's jump-off years	Period of evaluation	Variables
Australia	Adam 1992	1978–1989	1978–1990	TFR, CDR, net immigration, age structure
Austria	Hanika 1993	1977–1989	1977–1991	births, deaths, TFR, life expectancy by sex, total population, age groups 0–14, 15–59, 60+
Canada	Preston 1974	1950, 1954, 1960	1960–1968	life tables
	George and Nault 1991	1941–1983	1941–1990	total population, age structure, TFR, life expectancy, int. migration
	Statistics Canada 1994	1972, 1976, 1983	1972–1993	total population size
CSSR	Keilman and Kucera 1991	1950–1980	1950–1986	total population, age/sex structure, births, deaths
Denmark	Hansen 1993	1970–1990	1970–1991	births
	Spøhr 1995	1970, 1974, 1980, 1985, 1990	1975–1995	age/sex structure
	Spøhr 1995	1984–1994	1984–1994	births
	Spøhr 1995	1974, 1978, 1980, 1984–1986, 1988–1994	1974–1994	deaths
England	Field 1990	1949–1983	1954–1988	age structure (England and Wales)
		1974–1985	1974–1989	TFR (U.K.)
	Shaw 1994	1971–1992	1971–1992	total population, births, deaths, net migration, age structure (U.K.)
	Van Poppel and De Beer 1993	1970, 1977	1990, 1991	age/sex structure of the elderly (U.K.)
FRG	Bretz 1986	1952–1982	1963–1986	total population
	Van Poppel and De Beer 1993	1965, 1972	1990	age structure of the elderly
Finland	Hämäläinen 1987	1981, 1984	1981–1985	age structure (national), total population (municipal)
France	Vallin 1989	1964–1985	1964–1988	life expectancy by sex
Israel	Sabatello 1988	1965, 1970, 1975	1985	population 75+

Table 1. (continued)

	Reference	Forecast's jump-off years	Period of evaluation	Variables
Japan	Preston 1974	1949, 1954	1966	life tables
	Feeney 1990	1976, 1981, 1986	1976–1985	age- and sex-specific death rates between ages 40 and 79
The Netherlands	Keilman 1990	1950–1980	1950–1986	total population, growth rate, births, CBR, deaths, CDR, migration, age/sex structure, marital TFR (F1950, F1965, F1970), life expectancy by sex
	Cruijsen and Zakee 1991	1980–1989	1980–1991	total population, age structure, births, deaths, immigration, emigration, TFR, median age at childbearing, life expectancy by sex
	Van Poppel and De Beer 1993	1965–1980	1990	age/sex structure of the elderly
	De Jong 1995	1965–1994	1995	total population, broad age groups
	De Jong 1995	1975–1994	1975–1994	TFR, life expectancy by sex, net immigration
New Zealand	Preston 1974	1946, 1953, 1963	1965–1968	life tables
Norway	Brunborg 1984	1969–1982	1969–1983	births, population 80+
	Rideng 1988	1982	1984	total population by county, age structure for three municipalities
	Texmon and Keilman 1990	1969–1987	1969–1989	total population (F1946–F1989), age/sex structure, births, CBR, TFR, deaths, CDR, life expectancy by sex, net migration
	Texmon 1992	1969–1987	1969–1989	total population (F1946–F1990), age/sex structure, births, CBR, TFR, deaths, CDR, life expectancy by sex, net migration
	Statistics Norway 1994	1969–1990	1969–1993	growth rate (F1969–F1985), age/sex structure
Sweden	SCB 1986, 1989, 1991	1973–1989	1973–1990	births, deaths, net migration, 3 age groups
	Van Poppel and De Beer 1993	1965, 1970, 1975	1990	age/sex structure of the elderly

Table 1. (continued)

	Reference	Forecast's jump-off years	Period of evaluation	Variables
U.S.A.	Preston 1974	1947, 1958, 1962	1967, 1969	life tables
	Ahlburg 1982	1948–1975	1953–1980	births
	USBC 1984, 1989, 1992	1957–1982	1957–1986	population growth rate
	Long 1987	1945–1982	1945–1986	population growth rate, TFR, age groups 15–19 and 60–65
	Long 1995	1945–1986	1945–1992	population growth rate, TFR (Y1945–Y1985), age groups 15–19 and 60–64, total population size for states after 5 and 15 years (F1965–F1986, Y1970–Y1991)
OECD	Murphy 1987	1951–1984	1956–1985	total population, population by country, population by sex, age/sex structure, age structure by country
World	Keyfitz 1981	1939–1968	1950–1980	growth rates for Canada, the U.S., Europe and USSR, 9 countries of Eastern Europe, all countries of the world
	Stoto 1983	1950–1965	1955–1975	population growth rate by region (Japan, Western Europe, Southern Europe, Eastern Europe, Northern Europe, USSR, North America, Australia, and New Zealand)

indicators, such as the Total Fertility Rate (TFR) or the life expectancy, have, however, been evaluated in some cases. But sometimes one has to resort to even cruder variables for the components of growth, for instance crude rates or even absolute numbers of births, deaths, or migrants. The sub-sections in this section reflect the increasing level of detail.

Table 1 summarizes evaluation studies for national population forecasts and projections made by statistical agencies in 16 industrialized countries. At the end of the table, three studies are mentioned that analyse forecast accuracy for groups of industrialized countries. Because the aim is to discover possible *regularities* in error structures, the table is restricted to those studies in which *more than one forecast* was evaluated.

Many authors compute mean errors of some sort, for instance mean percentage error (MPE), root mean squared percentage error (RMSPE), or mean absolute percentage error (MAPE). MAPE is the average error when the direction of the error is ignored. This provides a measure of *accuracy*. MPE takes the direction into account. It provides a measure of *bias*: a positive MPE indicates that forecasts tend to be too high on average, and a negative MPE reflects too low forecasts. RMSPE is similar to MAPE, but gives more weight to large errors. Sometimes the mean is taken over various forecasts, or over various years, or forecast durations – sometimes it is unclear from the text how the mean error has been computed. This may create some problems in comparing the results. When forecast results are size dependent (total population size, numbers of births or deaths), accuracy measures that are independent of scale should be used in international comparisons – hence our focus on *percentage* error measures. When forecast results are not size dependent (e.g., population growth rate) other measures such as the mean error or its standard deviation can be used. These rules of thumb apply to comparisons of errors over time, between countries etc. However, the value of a single error is difficult to judge in isolation. Whether a particular over- or underestimation is to be regarded as large or small depends on the user's *loss function* (Keyfitz 1985): what costs are connected to positive and negative errors of various magnitude? Not many users are aware of their loss function; moreover, population forecasts produced by statistical agencies are general purpose forecasts computed for various users, and different users will have (at least in theory) different loss functions.

2.1. Total population size

Total population size is easiest to examine, but it does not provide much insight. The reason is that many forecasts of total population size were extremely accurate in the past, while at the same time both fertility rates and mortality rates were severely overestimated in many industrialized countries.[2] Thus it is not surprising to note that in the countries for which we have information, average errors in total population size generally are less than a few per cent, even with a forecast horizon of 15 years or more. Murphy (1987, p. 17) observes an average (over time and countries) root mean squared percentage error (RMSPE) of 4.9 per cent in forecasts of OECD member states issued between 1956 and 1979, when compared with actual outcomes during the period 1956–1985. When one controls for the length of the forecast period, RMSPEs in post-war forecasts for a period of

[2] Up to the end of the 1980s, errors in migration parameters were usually insignificant for the accuracy of total population size. Moreover, recent drops in life expectancy in some countries in Central and Eastern Europe have not yet shown up in forecast evaluations.

up to 15 years go down to 2–3 per cent; see Keilman (1990, p. 79) for the Netherlands, and George and Nault (1991, p. 9) for Canada.

Sometimes the mean percentage error (MPE) is used to evaluate forecast errors – e.g., for OECD countries, the Netherlands, Canada, Czechoslovakia, and Norway. Because overestimations and underestimations partially compensate each other in the MPE, this measure shows even smaller values than the RMSPE. It should be noted that in individual cases, relatively large errors (up to 20 per cent) are observed, see for instance the errors of U.S. population forecasts produced between 1891 and 1972, as computed by Ascher (1978, p. 50), and those of Scotland and Northern Ireland made in 1971 and evaluated in 1991 (8–10 per cent, see Shaw 1994, p. 28).

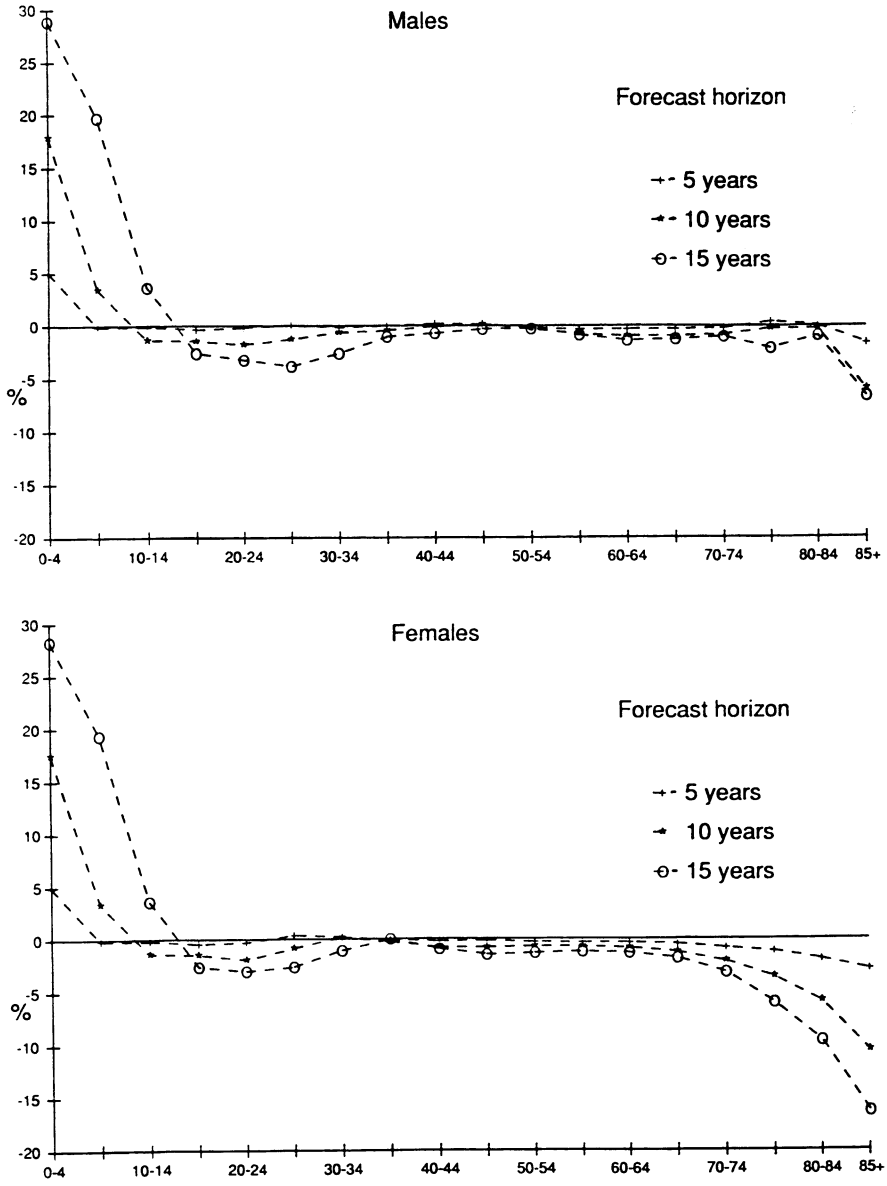
2.2. *Population growth rate*

Errors in population growth are frequently analysed by looking at the average annual growth rate of the population. This tradition was started by Keyfitz (1981) and Stoto (1983), who constructed confidence intervals surrounding population size point forecasts assuming a certain distribution of future errors in annual growth rates. For countries with low population growth, Keyfitz found a root mean squared error (RMSE) in the average annual growth rate of 0.29 percentage points. This is somewhat higher than the values for the Netherlands and Norway (0.20 percentage points in both cases). Long (1987, Table 1) shows that the RMSE in U.S. Bureau of the Census (USBC) growth rate forecasts varies between 0.2 and 1.1 percentage points for the forecasts produced between 1945 and 1970. Stoto computed a standard deviation (SD) of 0.50 percentage points for the error in annual growth rates for U.S. forecasts made after 1945. For Norway and the Netherlands, SD values of only 0.20 and 0.12 percentage points have been found, respectively. Compared with actual growth rates in industrialized countries, the errors reported here and elsewhere may be considered high.

Keyfitz's RMSE in the growth rate was only weakly dependent on forecast duration. But De Beer (1992) has argued that errors in growth rates must be autocorrelated. This has been confirmed for the Netherlands (De Beer 1992) and Norway (Statistics Norway 1994); in both cases it was found that a first order autoregressive model gives a good description of the error in the growth rate.

2.3. *Composition by age and sex*

The age structure summarizes the history of fertility, mortality, and migration at a certain point in time. This is the reason why countries that experienced a rapid fertility decline and a more or less steady improvement of mortality during the last few decades, display a typical pattern with strong positive errors (overestimations) at young ages, and moderate negative errors (underestimation) at more advanced ages. This pattern has been established for the United Kingdom, the Netherlands, Canada, Denmark, Norway, and Czechoslovakia (females only; numbers of elderly males were *over*estimated). An example is given in Figure 1 for the Netherlands. After 15 years, the average (over a number of post-World War II forecasts) MPE in the age group 0–4 years may be as high as 30 per cent (U.K., the Netherlands). For age groups over 85, females in particular display large errors; average MPEs of between −15 and −20 per cent (after 15 years of projection) have been found



Source: Keilman (1990).

Fig. 1. Mean percentage errors of forecasts of the age structure since 1950, the Netherlands

for Canada, Denmark, Norway, and the Netherlands. After 20 years the errors are even as large as 25–30 per cent for the over 85 in the U.K.

In sum, forecasters in a number of countries tended to overestimate young age groups and underestimate older age groups, particularly women. The reason is that they simply extrapolate birth and death rate series that exist at the time the forecasts are made. Official forecasters usually do not employ time series models or other formal methods for the analysis and extrapolation of fertility and mortality (Cruijsen and Keilman 1992, p. 10).

Such models may be useful in the short run, but they are less likely to provide accurate long-run forecasts (Land 1986). An important consequence of the mechanical or visual extrapolation methods that forecasters employ is that trend shifts in fertility or mortality patterns cannot be foreseen.[3] Hence we observe what has been called "assumption drag;" the continued use of assumptions long after their validity has been contradicted by the data (Ascher 1978, p. 53).

There are clear exceptions to the general error pattern caused by declining fertility and slowly improving mortality. In Finland, fertility increased in the early 1980s; life chances in that period improved less rapidly than expected. Therefore, for a large set of municipal forecasts made in 1981, Hämäläinen (1987, p. 8) found a pattern with large mean *negative* errors for low age groups, large mean *positive* errors for high age groups (80+), and only minor errors for intermediate ages. Positive errors have also been found for men aged 60+ in the 1965 forecast of Sweden, when numbers were compared with the observed age structure in 1990 (Van Poppel and De Beer 1993, Figure 7). Probably mortality assumptions for males have been too optimistic in the 1965 forecast for Sweden. In Section 2.5 we shall see that mortality forecasts made around 1960 in a number of industrialized countries have been overoptimistic for males aged 40–65.

An important indicator that summarizes a population's age structure is the dependency ratio, which relates the number of persons under 15 (young-age dependency ratio) and 65 and over (old-age dependency ratio) to the number aged 15–64 (or slightly different dividing lines between young, adult, and elderly age brackets). Surprisingly, no evaluation of the forecast accuracy of these dependency ratios has been found in the literature.[4] Yet these measures should receive more attention in accuracy studies, given their relevance for policy purposes.[5] We will present some new results for Austria, Norway, and the Netherlands.

Although Figure 1 suggests large overestimations of young age groups and large underestimations for the elderly, these errors are *relative to the actual value of each age group.* The dependency ratio, however, relates each of the two age groups to the (much larger) population of working ages, and the forecast accuracy of the dependency ratios may well be fairly high. Indeed, the figures that we computed on the basis of the data for Austria presented by Hanika (age groups 0–14 and 60+, as a percentage of age group 15–59) confirm this, see Figure 2. The downward slope in the young-age dependency ratio (YADR) has been forecasted accurately between 1977 and the mid-1980s (Figure 2a). However, the forecasts made between 1977 and 1983 suggested a weak increase during the second half of the 1980s, whereas in reality the YADR stabilized. The forecasted increase was caused by too high fertility assumptions (Hanika 1993, p. 15). The old-age dependency ratio (OADR) in Austria fell with two percentage points between 1977 and 1980. This was due to the decreasing size of the cohorts born around the end of the first world war. This trend in the OADR was accurately predicted by the forecasters in 1977. However, they did not foresee a slow increase taking place during the 1980s – indeed, mortality assumptions were too pessimistic (Hanika 1993, p. 15). The trend was picked up by the

[3] One may add that many non-theoretical formal models would not be able to predict such trend shifts either.
[4] Lee and Tuljapurkar (1994) investigate *ex-ante* errors in dependency ratios.
[5] It will be clear that the dependency ratios only indicate relative size in age groups. Data on income, labour market participation, social security etc. would be necessary to give the planner a more precise idea about relative sizes of subgroups giving or receiving support. Socioeconomic forecasts including such data have been made, but their evaluation is beyond the scope of this article.

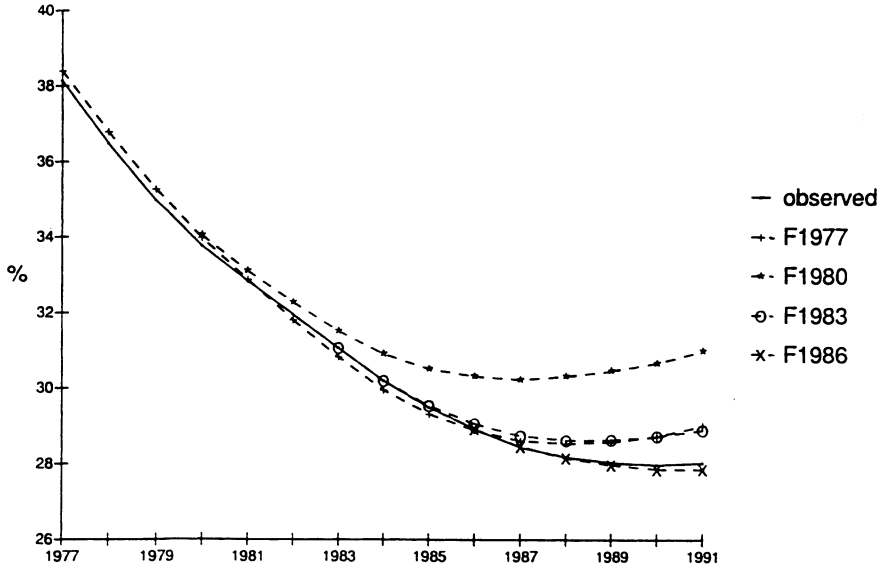
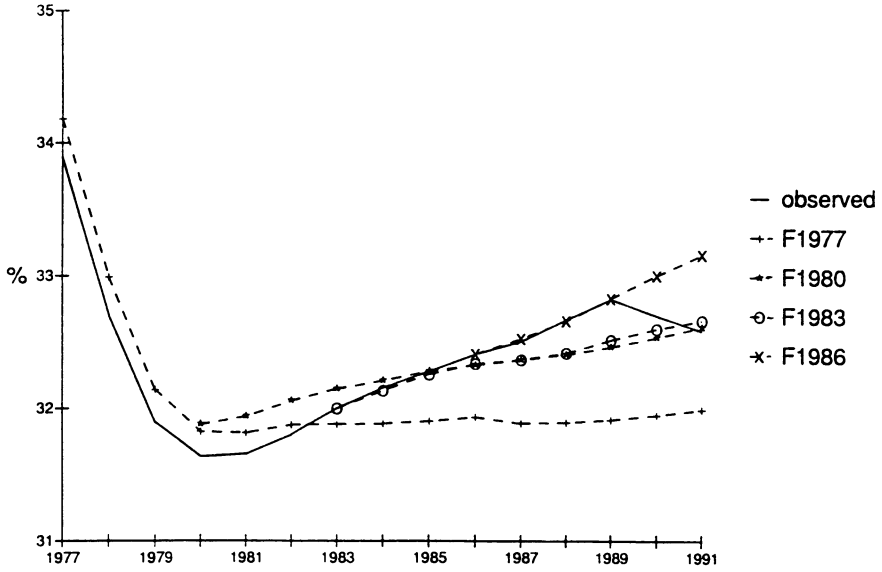


Fig. 2a. Young age dependency ratio, forecasts and observations, Austria

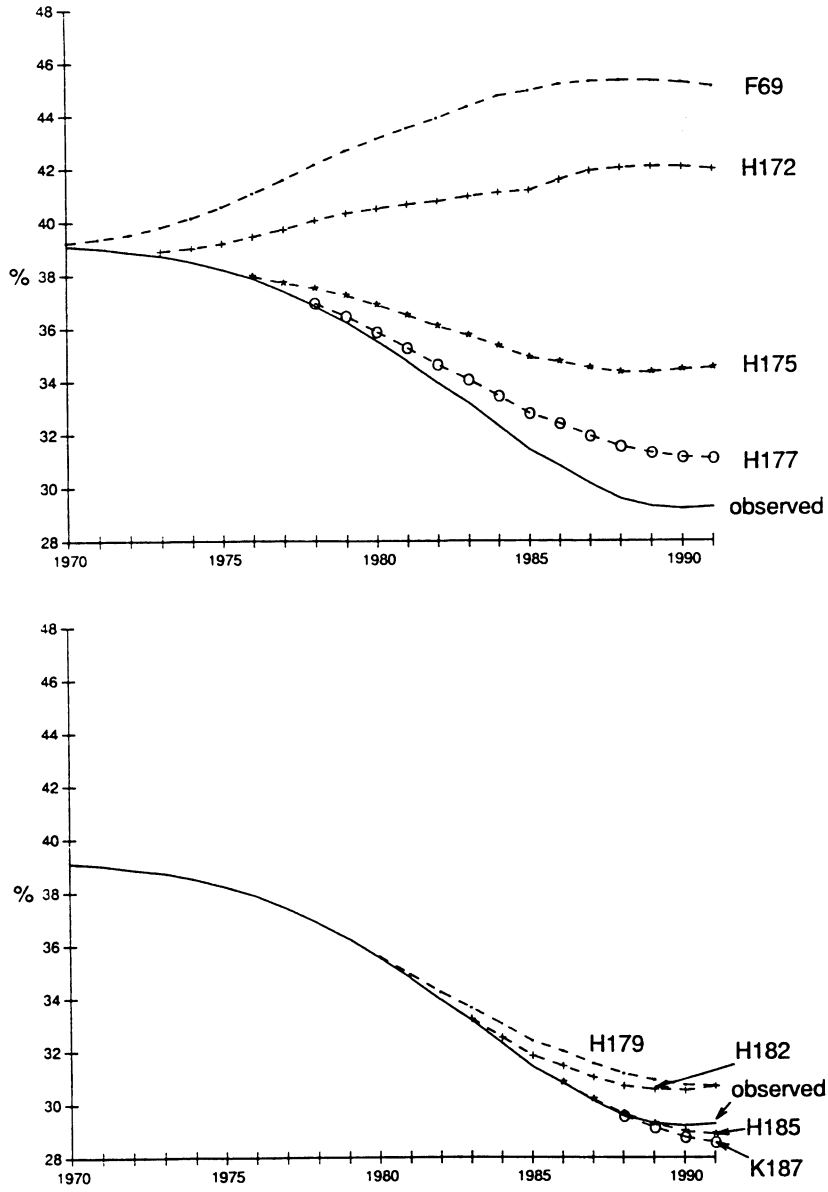


Source: Data published by Hanika (1993).

Fig. 2b. Old age dependency ratio, forecasts and observations, Austria

forecasters of 1980, 1983, and 1986, and the errors in OADR are half a percentage point or less. The unforeseen fall in the observed OADR after 1989 is explained by unexpected high immigration from East-European countries.

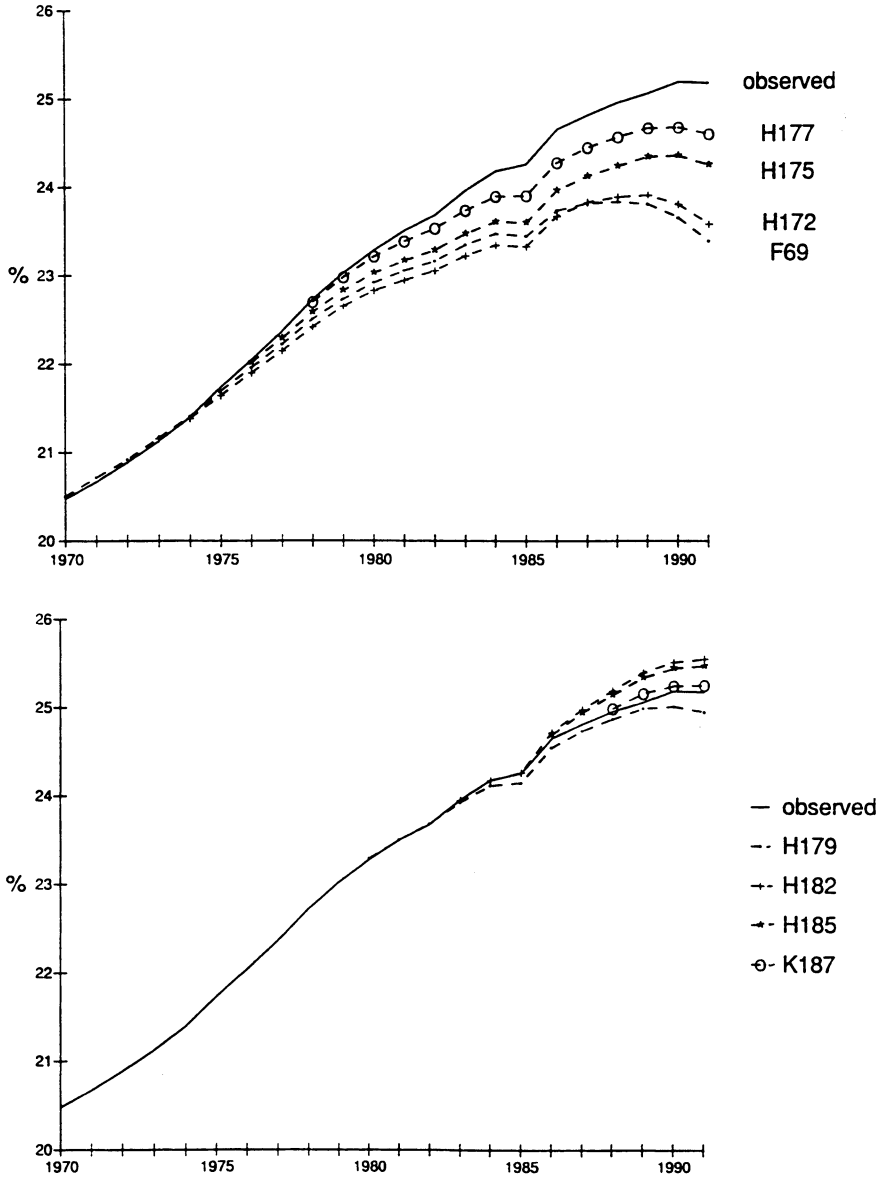
Figure 3 shows the YADR and the OADR for Norway, computed on the basis of age groups 0–14, 15–64 and 65+. The YADR (Figure 3a) exhibits the same pattern as the Austrian one, at least when restricting oneself to the forecasts made in 1977 or later. Those made in the period 1969–1975 show far too high values for YADR, caused by large



Source: Data collected by Texmon (1992).

Fig. 3a. Young age dependency ratio, forecasts and observations, Norway

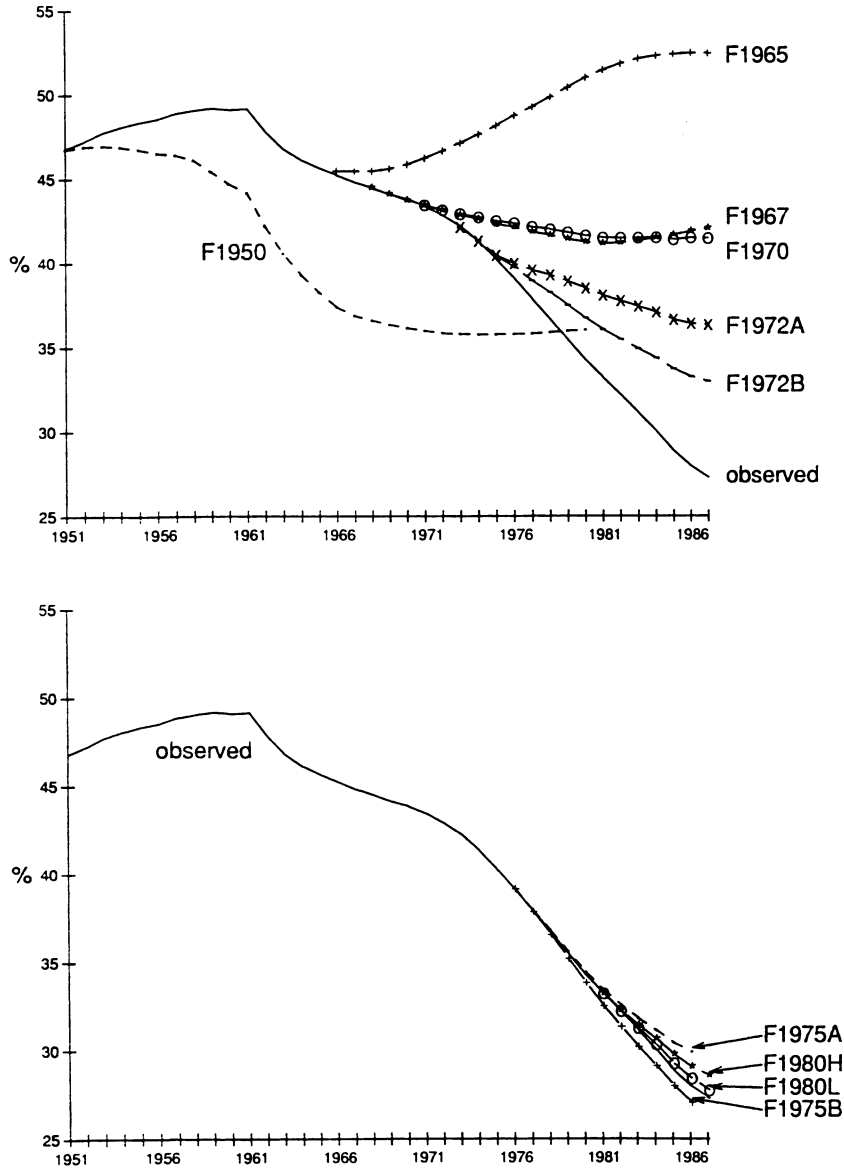
overestimations in the birth rates. The OADR (Figure 3b) shows consistent underestimations for the forecasts produced during the 1970s, as a consequence of too pessimistic assumptions regarding mortality. Yet the deviations are limited to two percentage points at most. The dip in both observed and forecasted OADRs in 1985 is caused by birth cohort 1920, which was exceptionally large, partly due to the marriage boom in the year 1918. Between 1985 and 1986, birth cohort 1920 disappeared from the denominator of the OADR, and entered the numerator.



Source: Data collected by Texmon (1992).

Fig. 3b. Old age dependency ratio, forecasts and observations, Norway

The performance of the dependency ratio forecasts for the Netherlands is illustrated in Figure 4. Forecasts made since 1965 show the same strong overestimations in YADR as those for Norway after 1969 (note that the scales of Figures 3a and 4a are different). The trend shift in 1961 is explained by the large cohorts born in 1946 and 1947 who reached age 15 in that and the following year. The OADR has been forecasted quite accurately, except in 1950, when mortality assumptions were far too pessimistic. The drop by one percentage point between 1978 and 1979 is due to irregularities in numbers of live births around 1914.



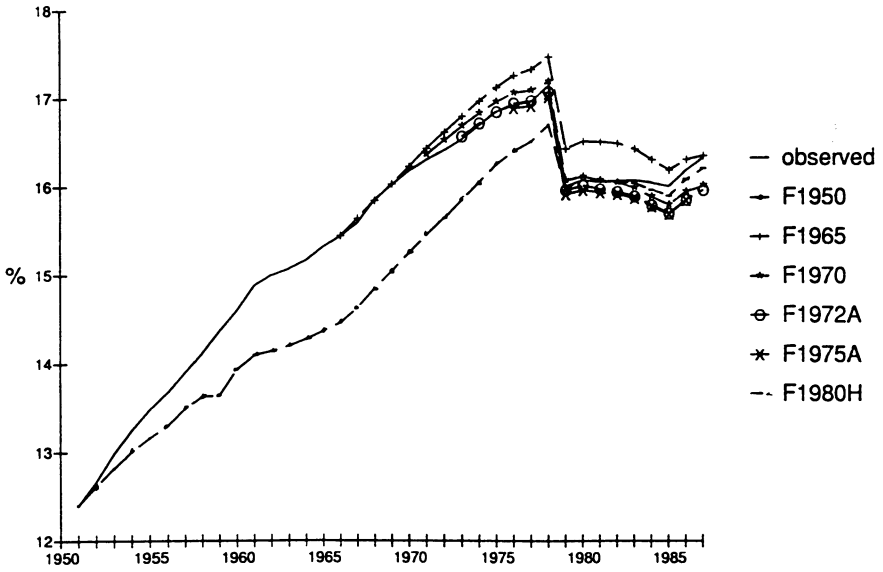
Source: Data collected by Keilman (1990).

Fig. 4a. Young age dependency ratio, forecasts and observations, the Netherlands

2.4. *Number of births, crude birth rate, TFR*

We have accuracy results for the number of forecasted births in Austria, the former CSSR, Denmark, the Netherlands, Norway, and the U.S. The studies by Shaw for the U.K. and Statistics Sweden only include graphs, and numerical conclusions are difficult to draw.

The range of percentage errors in births forecasts is quite wide: from an underestimation of 14–18 per cent (for the CSSR, Denmark, and the Netherlands) to an overestimation of 38–51 per cent (Denmark, the Netherlands, and Norway). The high positive errors in the



Source: Data collected by Keilman (1990).

Fig. 4b. Old age dependency ratio, forecasts and observations, the Netherlands

latter three countries all occur in the forecasts produced around 1970, when the steep fall in fertility was not foreseen. But the strong negative errors are spread over the post-war period: the 1970 forecast in the year 1980 in the CSSR (−16.8 per cent), the 1983 forecast in the year 1991 in Denmark (−17.8 per cent), and the 1950 forecast in the year 1965 in the Netherlands (−14.3 per cent). The recent underestimation in Denmark may be due to unforeseen high birth rates for older women. The downward trend in birth rates observed during the 1970s reversed in 1983, when fertility started to rise again. To a large extent this was due to the fact that women now had the births that had been postponed in earlier years. Indeed, birth rates for women between 25 and 35 years of age did not fall any longer, but started to increase in the mid-1980s. This catching-up effect was only partially foreseen in Danish forecasts produced after the mid-1980s.

The Crude Birth Rate (CBR) has been evaluated for the Netherlands and Norway. Its error patterns are very close to those of the number of births. Hence we will not discuss the CBR further here.

The accuracy of the Total Fertility Rate (TFR) extrapolations has been analysed for Australia, Austria, Canada, the United Kingdom, the Netherlands, and Norway. Error ranges are even wider than those for absolute numbers of births: from large underestimations (−33 per cent and −13 per cent for Canadian forecasts produced in the 1940s and Dutch forecasts made in 1950) to overestimations of between 44 per cent (Norway, F1969) and 79 per cent (the Netherlands, F1965). The forecasts made in the 1940s in Canada and the 1950-based forecast for the Netherlands did not foresee the baby boom during the 1950s and early 1960s. And when birth rates started to fall around 1970, this was not anticipated by official forecasters. The drop in the TFR was steeper than that in the absolute number of births, because the fall in the birth rate during the late 1960s

and the 1970s was partially compensated by increasing cohort sizes of the mothers, who themselves were born in the period 1940–1960. This resulted in somewhat larger errors for the TFR than for the number of births (the difference between error in TFR and that in births is six percentage points for the 1969-based forecast of Norway and eight percentage points for the 1965 forecast of the Netherlands). Not only Denmark, but also a number of other Western countries experienced an increase in their period Total Fertility Rates in the second half of the 1980s, caused by childbearing at higher ages after an initial period of postponement. This phenomenon has led to small underestimations (5–10 percentage points) of the TFR for recent Dutch and Norwegian forecasts.

2.5. *Number of deaths, crude death rate, life expectancy*

The errors in numbers of deaths range from −19 per cent (the former CSSR, F1965) to 17 per cent (Austria, F1977). Errors in crude death rates (CDR) are of comparable magnitude for those cases for which we have data. These error levels are much more modest than those for numbers of births. There are two reasons. First, forecast errors may be assumed to be relatively high when *behaviourally determined demographic variables* are concerned: migration, nuptiality, fertility and the like (Bartlema 1987, p. 21). Mortality, in which there is no or little choice involved, would lead to smaller forecast errors. Second, migration, nuptiality, and fertility are demographic phenomena that can be described by two aspects: their level (quantum) and their timing (tempo). How often does an individual migrate? How many children does he or she have? These questions are related to the *level* or *quantum* of the phenomena in question. Their *timing* or *tempo* is expressed by the age at which the corresponding events take place (or, more general, by the duration since some initial event). But for mortality, only questions related to timing are relevant. Since everyone dies, its level is 100 per cent by definition. This should make mortality easier to foresee than fertility, nuptiality, migration, and the like. Indeed, errors in Dutch mortality forecasts turn out to be only 20 to 45 per cent of those in comparable fertility forecasts. In many cases the ratio is 1 : 3 (Keilman 1990, p. 185).[6] In many industrialized countries the population growth rate is determined almost entirely by the (difference between the) crude birth rate and the crude death rate. Hence a 1 : 3 ratio of mortality errors to fertility errors would imply in general that one-fourth of the error in the growth rate would be due to mortality errors – provided one looks at *absolute* errors in these components, which cannot compensate each other in the growth rate error.

The life expectancy at birth is often used as a summary indicator for a whole range of age-specific death rates. Errors in this indicator are much lower than those in the number of deaths or the CDR. The underestimations and overestimations of the life expectancy are typically restricted to a few per cent of the observed values only. The reason for these small errors is that death rates at high ages contribute little to the life expectancy at birth,

[6] There is only one study into the timing aspect of *fertility*. Cruijsen and Zakee (1991, p. 36) evaluated four Dutch forecasts, made in 1980, 1984, 1986, and 1988, and compared forecast results with observed data for the period 1980–1990. Errors in the *median age at childbearing* were restricted to less than four per cent in absolute value. They show a more or less linear increase as a function of forecast duration. Errors in the TFR were much more irregular, but they were restricted to twelve per cent of the observed values. The ratio between errors in tempo and those in quantum is between 1 : 3 and 1 : 5.

but they have a large effect on the number of deaths and the CDR.[7] Thus it would be more useful to evaluate life expectancy at some advanced age (65, say) as a summary indicator for age-specific death rates. However no study that we are aware of has been concerned with this type of indicator.

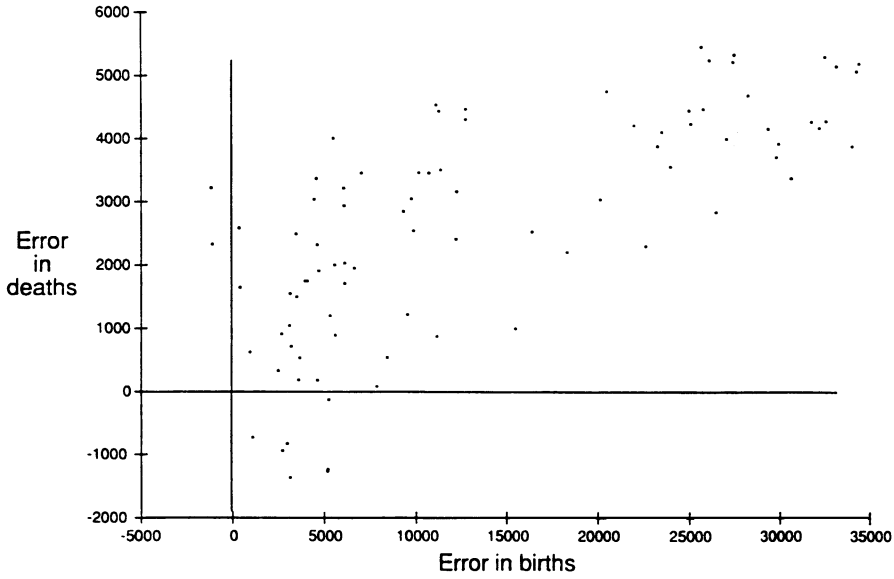
Concerning age-specific death rates, a detailed international comparison was carried out by Preston (1974). He evaluated official mortality forecasts made around 1950 and 1960 in Australia, Canada, Japan, New Zealand, and the United States. On the whole, Preston concludes, that for most groups defined by age and sex, mortality declined faster than was foreseen in the forecasts, in particular for females. This leads to large errors for females when mortality was assumed constant. Except for Japan, every forecast was over-optimistic for males in the age-range 40–65: future male mortality was underestimated for these ages in every one of the forecasts that Preston considered. His analysis reveals that resulting forecast errors are almost certainly attributable to unforeseen increases in mortality from neoplasms, cardiovascular diseases, and motor vehicle accidents.

3. Are Errors in Births and Deaths Independent? Evidence From Norway and the Netherlands

Alho and Spencer (1991, 1995) and Lee and Tuljapurkar (1994) construct stochastic models for the *ex-ante* specification of forecast errors and assume that the cross-correlation between fertility, mortality, and migration processes can be ignored. When the disturbances in these processes are not associated, this simplifies the model considerably. Lee and Tuljapurkar (1994, p. 1182) find only a weak correlation between errors in births and errors in deaths in the U.S. These errors are calculated as the residuals in time series models for fertility and mortality. In this section we shall take a different perspective and analyse a possible association between *ex-post* errors in historical forecasts of births and deaths. If such a cross-correlation is small, this further supports Lee and Tuljapurkars conclusion that it can be ignored, which leads to a considerable simplification of their and similar models.

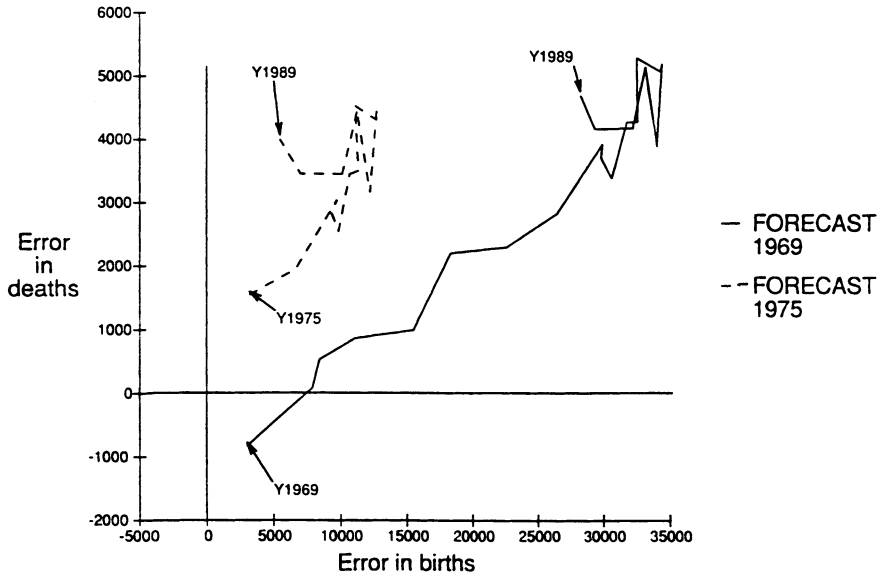
In general, there are various reasons why correlation could occur *across components.* For instance, in third-world countries, when infant mortality is high, this may lead to high fertility levels. A second reason may be unexpected economic shocks or variations in weather, which may affect both mortality and fecundity (Lee and Tuljapurkar 1994, p. 1182). In industrialized countries, one may speculate that large immigration numbers go together with high fertility, due to relatively high fertility among immigrant women. But it is less clear why cross-correlation could occur in the *errors* of components. One possible reason is the assumption drag that is commonly observed in population forecasts, cf., Section 2.3. If both fertility and mortality show new trends that are not picked up by the forecaster, this will lead to consistent positive or negative errors in both components, and hence to a positive or negative correlation. Signs of an assumption drag are clearly visible in many historical births and deaths forecasts: in spite of a rapid fall in fertility after the

[7] Another reason for the relatively low error values for the life expectancy is the fact that the latter indicator is not very sensitive to deviations in the age specific death rates (Keyfitz 1985, pp. 62–68). In Western countries for instance, an error in all death rates of 50 per cent of their respective values would result in an error in the life expectancy of only ten per cent.



Source: Data collected by Texmon (1992).

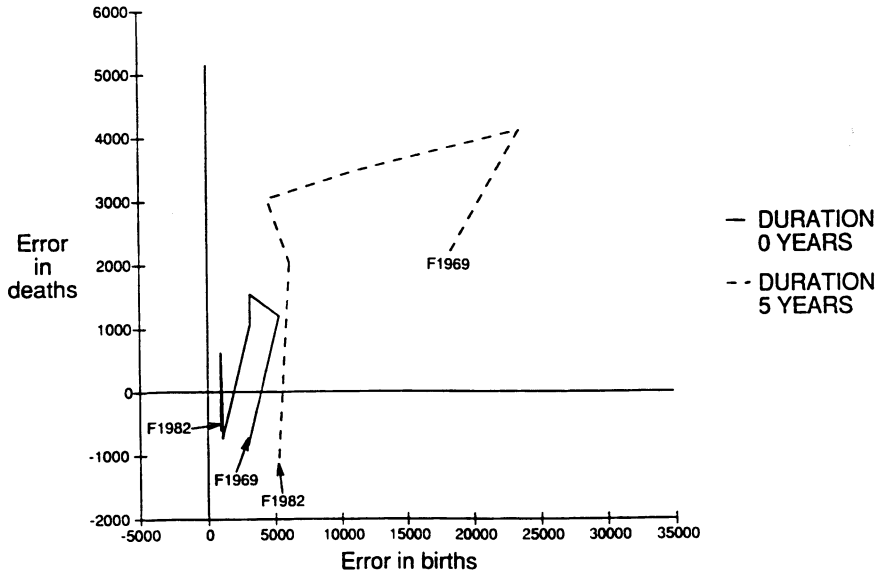
Fig. 5. Errors in births and deaths, forecasts 1969–82, calendar years 1969–89, Norway



Source: Data collected by Texmon (1992).

Fig. 6. Errors in births and deaths, forecasts made in 1969 and 1975, calendar years 1969–89, Norway[1]

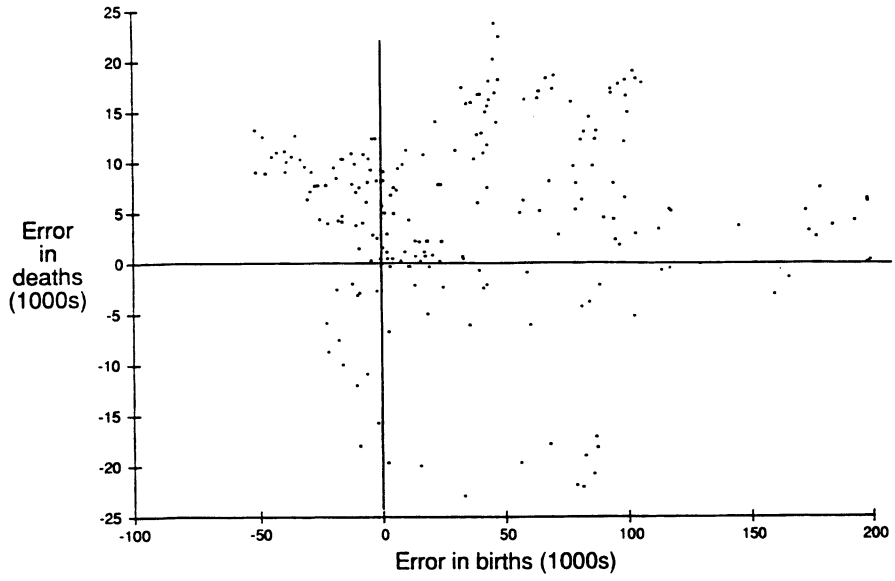
Note 1: "Y1969," "Y1975," and "Y1989" indicate observations for the calendar years 1969, 1975, and 1989, respectively.



Source: Data collected by Texmon (1992).

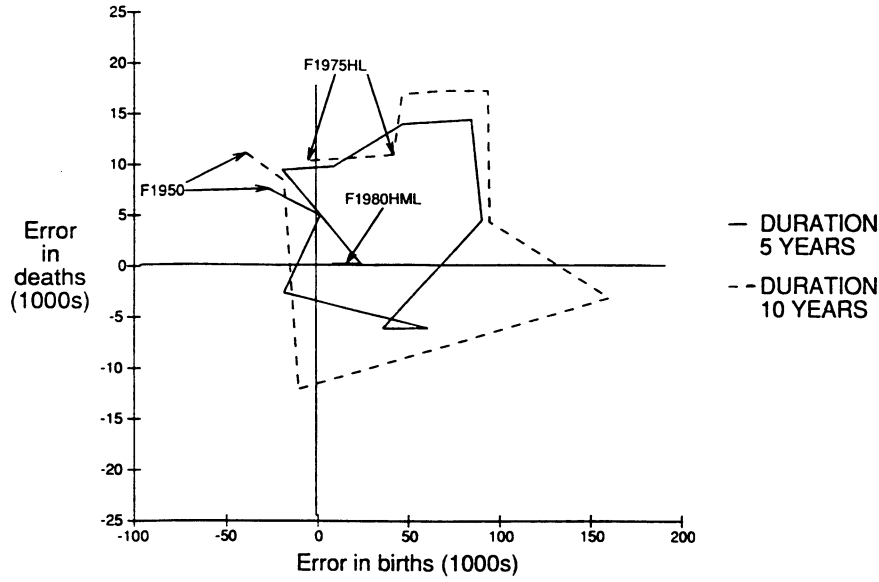
Fig. 7. Errors in births and deaths, forecasts 1969–82, durations 0 and 5 years, Norway[1]

Note 1: "F1969" and "F1982" indicate observations for the forecasts with base year 1969 and 1982, respectively.



Source: Data collected by Keilman (1990).

Fig. 8. Errors in births and deaths, forecasts 1950–80, calendar years 1950–86, the Netherlands



Source: Data collected by Keilman (1990).

Fig. 9. Errors in births and deaths, forecasts 1950–80, durations 5 and 10 years, the Netherlands[1]

Note 1: "F1950" indicates observations for the 1950-based forecast. The two points indicated by "F1975HL" at duration ten years represent errors observed for births and deaths at that duration according to the high and the low variant of that forecast in the year 1985. Similarly, the three points indicated by "F1980HML" at duration five years represent errors according to the high, the medium, and the low variant of that forecast in the year 1985.

mid-1960s, and an improvement in mortality around 1970, many of the forecasts referred to in Table 1 are based upon an extrapolation of the old patterns.

Ideally, one would like to have errors in age-specific rates for each of the components in order to investigate the cross-correlation structures. But information at so detailed a level is not readily available for historical forecasts. Therefore, as a first step, we have looked at errors in numbers of births and numbers of deaths. We have used the data set for Norway constructed by Texmon (1992), and one for the Netherlands (Keilman 1990), see also Table 1. For Norway, the data cover the forecasts made between 1969 and 1987, and calendar years 1969–89. In order to have more than only a few observations for each forecast, the analysis is restricted to the forecasts made between 1969 and 1982. For the Netherlands, forecasts made between 1950 and 1980 are included, with errors for the years 1950–86. Simple scattergrams have been used to investigate the cross-correlations.

Figure 5 for Norway illustrates errors in births and deaths for all forecasts with jump-off years between 1969 and 1982. The observations cover the period 1970–1989. Each dot represents the error in numbers of births and deaths for a particular forecast in a particular year. The cloud appears to display a weak positive correlation, with errors in births increasing somewhat faster than those in deaths. When we look at separate forecasts, the correlation is particularly strong for the 1969 forecast, see Figure 6. Between the years 1969 and 1982 the relationship is almost linear. However, this is largely explained by a duration effect: forecast errors tend to grow with increasing length of the forecast period (Ascher 1978; Stoto 1983; Armstrong 1985; Smith 1987; see also estimation results for

duration effects in Section 4). Therefore, as soon as one controls for the duration effect, the positive correlation between errors in births and deaths disappears. For instance, Figure 7 gives errors at duration 0 years (i.e., errors made during the jump-off year – jump-off populations were defined as of 1 January) and five years. For both durations, we have observed errors connected to the forecasts of 1969 (indicated by F1969), 1972, 1975, 1977, 1979, and 1982 (indicated by F1982). Given a certain level of the error in the births forecast, there is little we can say about the error in the deaths forecast, and vice versa.

Data for the Netherlands, displayed in Figure 8, show a weaker correlation than the Norwegian data in Figure 5. The forecasts of 1965 and 1970 appear to possess a certain degree of positive correlation (not shown here). But also in this case the relationship between births errors and deaths errors disappears when we control for forecast duration, see Figure 9 for durations of five and ten years.

Although the number of data points in both data sets is not very large, we find little support for a correlation between errors in births and deaths forecasts, *provided that we control for forecast duration.* Therefore it is contended that a similar analysis on the basis of age-specific rates, instead of absolute numbers of events, would not reveal a strong cross-correlation either.

4. Did Forecast Accuracy in Norway and the Netherlands Improve in Recent Decades?

The findings presented in Section 2 relate to *old* forecasts. These results may be used to assess the accuracy of *current* forecasts provided that one is willing to assume that current forecasts have error levels similar to those of the older ones. However, an obvious assumption would be that errors have diminished over time: new data have become available, more refined techniques have been developed, and new information technology has facilitated rather detailed analysis. But is it really the case that old forecasts exhibit larger errors than more recent ones? This question will be investigated for two countries for which we have detailed information. We have used the data sets for Norway and the Netherlands introduced in Section 3.

A naive approach would be to compute some mean error for each forecast and next see whether more recent forecasts have smaller mean errors than the older ones. But this approach is not without problems. Take for example the forecasts of Norway with jump-off years 1969 and 1972. One could compare the ex-post observed error in births arising in the 1969 forecast for, say, the year 1974, with the ex-post error of the 1972 forecast for the same year. However, the 1969 forecast was five years old in 1974, and the 1972 forecast only two years. Because of the duration effect discussed in the previous section, the errors of the two forecasts as observed in 1974 cannot be compared directly. On the other hand, when the disturbing effect of the length of the forecast period is held constant, one would have to compare the ex-post errors that are observed in different calendar years, in which the forecastability of births may not be the same. Therefore, when analysing the ex-post errors of a series of forecasts, one should separate, for each forecast, the effects of forecast duration and calendar year.

In this section we shall investigate ex-post errors in Norwegian and Dutch fertility and mortality forecasts by means of a multivariate model that distinguishes between period

(calendar year) effects, duration (length of forecast period) effects and jump-off year effects. The jump-off year effect of a particular forecast would tell us what the "mean" level of error is which is connected with that forecast. It can only be estimated accurately if confounding effects of period and of duration are removed from the observed forecast errors. By comparing jump-off year effects for subsequent forecasts, we would be able to investigate a possible improvement of forecast quality over time.

Modelling and estimation aspects have been discussed in detail elsewhere (Keilman 1990, 1991), and therefore these technical aspects will be dealt with only briefly here. The model distinguishes between period (calendar year) effects, duration (length of forecast period) effects, and jump-off year effects in the errors observed for a series of forecasts. This model is, in fact, equivalent to the well-known age-period-cohort model (APC model).

The *cohort effect* of the usual APC model is interpreted here as the contribution to the forecast error of the particular forecast (jump-off year). This *forecast effect* is seen as an indicator that expresses in one number the various aspects that are specific for the forecast under consideration, for instance the composition of the forecast team, or the methodology used.

The *age effect* used in APC models here reflects the error contribution of the length of the forecast period. This so-called *duration effect* expresses the increase in uncertainty as the forecast period grows longer. The longer the forecast period, the greater the probability that external factors, which are assumed constant in the forecast, will change.

The *period effect* of the present model corresponds to the period effect of APC models. It quantifies a set of factors that expresses social and demographic conditions observed during a certain period in time, and their effect on forecast errors, independent of the circumstances under which the forecast was produced (these are summarized in the forecast effect) or the length of the forecast period (captured by the duration effect).

In its most simple version, the separation model takes on the following general form.

$$|E(i,j,k)| = \exp\{F(i) + P(j) + D(k) + u(i,j,k)\} \tag{1}$$

Equation (1) explains the absolute error for the forecast with jump-off year i, as observed in calendar year (or period) j, for a duration of k years, as an exponential product of a forecast effect $F(i)$, a period effect $P(j)$, a duration effect $D(k)$ and a residual $u(i,j,k)$. The error is defined as the absolute value of forecast minus observed variable. In the model it is assumed that all errors that are observed for one particular forecast, e.g., the one with jump-off year 1972, have the same forecast effect (to be denoted by $F(1972)$ in this example), irrespective of the forecast's duration, or the year for which these errors are observed. Similarly, all errors observed in a certain calendar year have the same period effect $P(j)$, regardless the jump-off year of their respective forecasts, while the duration effect for a duration of k years ($D(k)$) applies to all errors of forecasts that are k years "old." A positive effect estimate implies, when exponentiated, a contribution to the error that is higher than some (unknown) "mean" error or "baseline" error, and a negative effect estimate implies a contribution lower than the mean (baseline) error.[8]

[8] The usual warning in regression models with categorical variables applies here: estimated effects as such are without meaning, only the *difference* between two effects can be interpreted.

Model (1) suffers from perfect multicollinearity between forecast jump-off year (i), period (j), and duration (k): given two of these variables, the third is completely determined by the relationship $k = j - i$. In APC-models this is known as the identification problem. We have overcome this problem by parameterizing the duration effect (see below), and by aggregating (most of) the period effects in five-year intervals. This breaks the perfect identity $k = j - i$.

Model (1) contains many effects only. It can be extended with interaction effects, when necessary. For instance, when the forecast with jump-off year i is assumed to exhibit period effects for period j that are different from those in forecasts with other jump-off years, an interaction term $FP(i,j)$ may be added to model (1). Interaction effects will turn out to be necessary in a few cases.

Special attention has to be given to the issue of forecast *variants*. The two agencies formulated more than one fertility variant for the forecasts of 1972 and later. Therefore an additional term $V(m)$ has been included in expression (1), which can be interpreted as variant effect.

The duration effects will be parametrized as $D\sqrt{k}$, with D a coefficient to be estimated from the data and k is defined as calendar year of observation minus forecast's jump-off year. Compared to a linear form or an exponential form for duration dependence, a square-root form has been found to work well in earlier analyses (Keilman 1990).[9]

Tables 2 and 3 show regression results for Norway and the Netherlands, respectively. The separation model has been applied to errors in numbers of live births and deaths.

Norwegian births forecasts made between 1969 and 1982 show a steady improvement in accuracy. Whereas the forecast with jump-off year 1969 still had an error that was $\exp(1.639) = 5.15$ times as large as the average error, the one with jump-off year 1982 had births errors that were only 0.58 times the average. Most of the improvement took place between 1969 and 1977; the forecasts made in the period 1977–1985 show little variation in the estimated main effects. However, the 1982 forecast and the 1985 forecast had births errors that increased more sharply than on average. Hence the estimated interaction between forecast effect and duration effect is positive for both forecasts. The amount of variance explained by the model is not impressive (61 per cent). By adding more interaction effects this could be improved, probably, but experience with this model fitted to Dutch data (a simple linear model for signed errors) suggests that the differences in forecast effects for subsequent forecasts do not change much, once main effects have been included in the model (Keilman 1990).

The method used for formulating fertility assumptions changed in 1975: the 1975 forecast was the first one in which observed trends were explicitly and gradually extrapolated into the future. In 1969 and 1972, future birth rates were simply obtained by keeping recently observed rates constant (Texmon 1992, p. 290). To what extent did this change in forecast methodology improve the accuracy? The 1975 forecast did, indeed, have smaller errors (0.95 times the average) than the 1972 forecast (2.62 times the average). But the improvement in accuracy had started already in 1969, and it continued until 1977. Therefore it is unlikely that the change in methodology has caused a clear reduction in the errors in births. The high error for the 1987 forecast is explained by the fact that the

[9] Joop de Beer has pointed out to me that a square root form of duration dependence for the errors also arises when they are generated by a random walk process.

Table 2. Regression results separation model for errors in births and deaths, Norway

	Births			Deaths		
	Estimate	Exp.*	*t*-value of estimate	Estimate	Exp.*	*t*-value of estimate
Effects						
Jump-off years						
F(1969)	1.639	5.15	2.87	−1.628	0.20	−2.75
F(1972)	0.962	2.62	1.78	−0.944	0.39	−1.59
F(1975)	−0.050	0.95	−0.09	−0.727	0.48	−1.24
F(1977)	−0.531	0.59	−0.98	−0.942	0.39	−1.60
F(1979)	−0.478	0.62	−0.88	−1.138	0.32	−1.89
F(1982)	−0.539	0.58	−0.85	−1.868	0.15	−2.55
F(1985)	−0.405	0.67	−0.55	−1.001	0.37	−1.43
F(1987)	1.693	5.44	2.03	0.330	1.39	0.44
Periods						
P(1970–74)	0.142	1.15	0.27	−0.341	0.71	−0.59
P(1975–79)				−0.157	0.85	−0.26
P(1980–84)	0.247	1.28	0.58	−0.574	0.56	−0.81
P(1985–89)	−0.281	0.76	−0.54	−1.331	0.26	−1.66
P(1973)	−2.151	0.12	−3.35			
P(1975)	0.425	1.53	0.77			
P(1976)	0.353	1.42	0.65			
P(1977)	0.676	1.97	1.38			
P(1979)	0.287	1.33	0.62			
Durations						
D(1)				1.436	4.20	4.57
$\sqrt{k}$	0.488	1.63	2.74	1.023	2.78	5.74
Variants						
V(High)	0.503	1.65	1.51			
V(Constant)	−0.735	0.48	−1.57			
V(Low)	−0.523	0.59	−1.40			
Interactions						
k.F(1982)	0.235	1.26	2.42	0.102	1.11	1.21
k.F(1985)	0.512	1.67	2.67			
R^2(adj.)		0.61			0.73	

* Exponentiated value of estimated effect.

catching-up effect in fertility which started in Norway in the mid-1980s was not foreseen. This led to a strong underestimation in the number of births. On the other hand, the estimate is not very reliable as it is based on very few observations.

Period effects fluctuate rather strongly, in particular during the 1970s. This is the reason why separate effects have been estimated for the years 1973, 1975, 1976, 1977, and 1979. The period effects for 1973 and 1985–1989 happen to be negative (for the year 1973 the

effect is equal to $P(1970–74) + P(1973) = 0.142–2.151 = -2.009$), which implies lower errors for those years than on average.

The positive coefficient for $\sqrt{k}$, which is highly significant, implies a strong autocorrelation between errors. For instance, at durations one, four, and nine years, the errors are 1.63, $1.63^2 = 2.65$ and $1.63^3 = 4.32$ times the average error. A similarly strong autocorrelation occurs for errors in deaths, and for error data for the Netherlands, cf. below.

The error patterns for deaths are much more irregular than those for births. All forecasts but the 1987 forecast had errors in deaths below the average, but to varying degrees.[10] A monotone improvement in forecast accuracy is not visible in the data. The duration effect is much stronger than both forecast and period effects.

Births forecasts in the Netherlands show a very irregular error pattern, see Table 3. Forecasts made in 1965 and 1970 had large errors: 16.1 and 11.7 times the average error, respectively. In 1975, forecast assumptions were formulated, for the first time, on the basis of sociodemographic interpretations of observed trends, not only mechanical extrapolations. This seems to have improved the accuracy of the 1975 forecast, compared to that of forecasts made since 1965. The 1975 forecast also contains clear effects caused by the use of a high and a low variant of fertility: the interaction effect *k.F(1975).V(High)* indicates that in the high variant of that forecast, errors followed a linear upward pattern as a function of duration (k). This interaction effect alone leads to errors in the high variant of the 1975 forecast that are 15 per cent higher than the average error – for the low variant the errors are two per cent lower than the average. The margin between the high and the low variant was much larger than that of other forecasts (for which no significant variant effect could be estimated). It may be explained as a reaction to the relatively large errors in the fertility forecasts of 1965–1972.

Concerning Dutch deaths forecasts, the accuracy of subsequent forecasts shows no clear pattern. The 1965 forecast had remarkably low overall errors (48 per cent below the average), but this estimate concerns all observed errors until 1986. For the first few years, however, the forecasters noted (in 1970) that they had been too optimistic regarding male mortality rates in the 1965 forecast for ages up to 54 years, and for female rates even up to 69 years (Keilman 1990, p. 97). This is the main reason why the 1970 forecast had *increasing* mortality rates for some age groups which resulted in a slightly decreasing life expectancy for males and only a slow improvement for females – in reality, both sexes showed a considerable increase in life expectancy in the 1970s. This is another example of a reaction pattern displayed by forecasters who note the errors in previous forecast: as a result, errors change sign. Indeed, the errors in the 1965 forecast were no less than 48 per cent below the average error, as noted above, whereas those in the 1970 forecast were (main effect *plus* interaction effect) $\exp(-0.040 + 0.131) = 1.10$ times as large. The overestimations in the numbers of deaths continued for the forecasts of 1972 and 1975.

In sum, there is some tendency of an improvement in the accuracy of births forecasts for Norway between 1969 and 1977; after that errors stabilized at a level of 60 per cent of the average. For the Netherlands, birth forecasts became clearly more accurate between 1970 and 1975, but when we look at the longer period 1950–1980 there is no uniform

[10] The large but insignificant effect for the 1987 forecast is based upon only a few observations.

Table 3. Regression results separation model for errors in births and deaths, the Netherlands

	Births			Deaths		
	Estimate	Exp.*	*t*-value of estimate	Estimate	Exp.*	*t*-value of estimate
Effects						
Jump-off years						
F(1950)	0.811	2.25	1.66	0.378	1.46	0.58
F(1951)	−0.010	0.99	−0.02	0.955	2.60	1.18
F(1956)	0.628	1.87	1.46	0.742	2.10	1.32
F(1965)	2.781	16.14	8.06	−0.651	0.52	−1.50
F(1967)	2.020	7.54	5.91	−0.256	0.77	−0.63
F(1970)	2.459	11.69	7.54	−0.040	0.96	−0.09
F(1972)	1.994	7.34	6.85	1.477	4.38	4.42
F(1975)	0.364	1.44	0.96	1.327	3.77	4.37
F(1980)	1.297	3.66	5.50	−0.839	0.43	−3.50
Periods						
P(1951–55)	1.044	2.84	2.67	0.688	1.99	1.28
P(1956–60)	0.556	1.74	1.73	0.518	1.68	1.25
P(1961–65)	0.940	2.56	3.60	0.164	1.18	0.50
P(1966–70)				0.202	1.22	0.79
P(1971–75)	0.447	1.56	2.73			
P(1976–80)	0.679	1.97	4.72	0.080	1.08	0.55
P(1968)	−0.959	0.38	−2.95			
P(1970)	−1.122	0.33	−3.52			
P(1972)				−0.222	0.80	−0.71
P(1973)				−0.727	0.48	−2.60
P(1974)				0.271	1.31	1.00
P(1981)				−0.179	0.84	−0.78
P(1982)	0.371	1.45	1.68			
P(1983)	0.255	1.29	1.15			
P(1985)				−1.165	0.31	−4.81
P(1986)				−1.304	0.27	−5.25
Durations						
D(1)	−0.779	0.46	−3.47			
$\sqrt{k}$	0.572	1.77	6.55	0.470	1.60	4.31
Interactions						
k.F(1970)				0.131	1.14	3.42
k.F(1975).V(High)	0.140	1.15	2.63			
k.F(1975).V(Low)	−0.024	0.98	−0.46			
$\sqrt{k}$.F(1951)				−0.345	0.71	−2.70
R^2(adj.)		0.74			0.71	

* Exponentiated value of estimated effect.

improvement. For death forecasts the accuracy is more volatile in both countries, and one must conclude that no clear improvement has taken place.

5. How Can Ex-Post Errors Be Included in Current Forecast Reports?

Despite the fact that statistical agencies apply a strictly deterministic approach for their population forecasts, a number of instruments can be used, even in such an approach, to express forecast uncertainty somewhat more informally than by means of a stochastic model, and some of these instruments are actually employed by the agencies. First, ex-post errors observed for historical forecasts are a useful way of summarizing uncertainty for main forecast variables. For instance, the forecast reports published by Statistics Sweden in 1986, 1989, and 1991 contain graphs with observed and forecasted numbers of births, deaths and immigration.[11] Statistics Norway presented in its report on the 1993-based forecast a graph with mean percentage errors for forecasts of the age structure. The latter type of figure is of particular value to the user, who is interested in forecasted age structure, rather than numbers of births, deaths, or immigrants.

Second, these ex-post observed errors may be used to express the uncertainty of the *current* forecast. Population forecasts traditionally rely on the extrapolation of observed trends in fertility, mortality, and migration parameters. Error statistics may be extrapolated as well, at least in the short run. For instance, in a new forecast, the distance between the high and the low variant of a certain parameter in the first few years of the forecast can be set on the basis of ex-post forecast errors of the relevant parameter in historical forecasts (De Beer 1988). This is actually done, for the first few years of the extrapolation period, by Statistics Netherlands and Statistics Norway. For Norway, we know the standard deviation of the error in the life expectancy forecasts made between 1969 and 1990, thereby controlling for forecast duration (Texmon 1992, p. 306). In the 1993-based mortality forecast, the life expectancy for men and women was the key variable to be extrapolated. The life expectancy in the high variant for each of the first five years has been determined as the corresponding variable in the medium variant plus one standard deviation, and similarly for the low variant (Statistics Norway 1994, p. 23). Even more formal is fitting a time series model to observed forecast errors of some variable, and making a prediction of the confidence interval for that variable to be applied in the current forecast. For instance, De Beer (1992) found that an AR(1) model could be fitted to errors in the growth rate of total population size. The error in total population size equals the cumulated error in growth rates since the base year. De Beer found a simple expression for predictions of the confidence bounds around a given forecast of total population size (for example, the medium variant taken from an official forecast). This method has been used by Statistics Norway for assessing the probability connected to the interval defined by the high and the low variant: these two variants (formulated independently from the AR(1) model for the growth rate error) turned out to agree rather closely with the boundaries of a 67 per cent confidence interval. Also the high-low fan of forecasted total population size for the U.S.A. turns out to coincide quite well with a two-thirds confidence interval, although the latter interval has been predicted on the basis of an assumed *constant* error in the growth rate – not by means of a time series model (USBC 1989).

[11] The report on the 1994-based forecast of Statistics Sweden no longer contains such figures.

6. Conclusions

Users of statistics published by statistical agencies have the right to know how reliable the figures are. The quality of a population census is often evaluated by means of a post-enumeration survey. Likewise, the quality of a population *forecast* can be assessed. National statistical agencies in industrialized countries have produced population forecasts over a sufficiently long period, and therefore we can evaluate the accuracy of those old forecasts by comparing their results with ex-post observed actual trends. This indicates to what extent we can rely on forecasts that are *currently* produced – provided we are willing to assume that forecasting nowadays is as difficult as it was in the past. In case we are more optimistic, the historical errors would give us an upper bound to the expected errors in current forecasts.

Quality assessment is not the only reason why insight in forecast errors is useful. A second reason is that such errors provide the user with quantified insight into forecast uncertainty. Pending the development of operational stochastic models that would supply the user with confidence intervals, and not just point predictions, errors observed for old forecasts (as a second best) indicate how much flexibility has to be included in the planning process, at least concerning population numbers.

In this article we have given a broad overview of ex-post observed errors in historical population forecasts at the national level in industrialized countries. The literature review in Section 2 showed that:

- percentage errors in the age structure display a typical pattern in many countries, with strong overestimations for young age groups, and considerable underestimations for the elderly;
- the large errors found for both the young and the old after a forecast period of 15 years (up to +30 per cent for the age group 0–4, and −15 per cent or lower for women aged 85+ are not uncommon) suggest that those old forecasts supplied useful information perhaps up to 10–15 years ahead, but certainly not longer;
- errors in fertility are much higher than those in mortality – behaviourally determined variables are difficult to forecast;
- reported errors in the Total Fertility Rate range from −30 per cent to +80 per cent – this range is much wider than that for errors in absolute numbers of births, or the Crude Birth Rate;
- life expectancy at birth shows only minor forecast errors, in spite of considerable errors in numbers of deaths and Crude Death Rates – life expectancy at elderly ages will presumably show larger errors, but no analysis of the forecast accuracy of the latter indicator is known of;
- too little attention has been given to the evaluation of dependency ratio forecasts.

An analysis of two rather detailed data sets (Norway, forecasts made between 1969 and 1982; the Netherlands, forecasts made between 1950 and 1980) revealed no systematic correlation between errors in numbers of births and errors in numbers of deaths. Improvements in the accuracy in Norwegian births forecasts can be observed for forecasts made between 1969 and 1982, although no link could be established with changes in the method for the formulation of fertility assumptions. There was no improvement after 1982.

Norwegian deaths forecasts, and Dutch forecasts both for births and deaths, showed no systematic reduction over time in their errors. Thus one must be prepared for errors in future forecasts that are of similar magnitude as those in the old forecasts. Strong autocorrelations occur both in the births errors and the deaths errors for the two countries.

Several methods exist for communicating forecast uncertainty to the user. For instance, one can use simple graphical illustrations that show mean errors for historical forecasts around the age structure, or that compare observed and forecasted numbers of births, deaths and immigrants.

In general, errors in historical population forecasts at the national level seem to have been caused, to a large extent, by changes in real observed demographic variables. This may seem to be a tautology, as the forecast error in some variable is defined as the difference or the ratio of the forecasted and the ex-post observed value. But the point is that forecasted trends are smooth extrapolations of trends that were observed before the forecast was made, whereas reality may exhibit sudden trend shifts and other unexpected irregularities. These irregularities appear to dominate ex-post observed forecast errors.

A second important factor that explains errors in historical forecasts is the choice of assumptions made for the relevant parameters by the forecaster or the forecast team. Two possibilities have frequently been encountered: assumption drag, and over-reaction. An assumption drag seems to have been present among forecasters in the past, and probably it still is. New trends exhibited by important demographic variables, for instance the sharp decline in birth rates in the 1970s, or improved male life expectancies in the 1970s after a period of stagnation, are only picked up by the forecasters some ten years after these new trends have set in. Because no behavioural models of sufficient explanatory quality exist that can be used for extrapolating fertility and mortality, demographers must rely on observed trends, and any new trend in the time series must have been observed over a long enough period before an accurate extrapolation can be carried out. Indications of overreactions have been found for fertility and mortality in the Netherlands. A strong under- (over-)estimation of some variable in a previous forecast leads to too high (low) assumptions in the next forecast, which then results in an over- (under-)estimation of that variable. It is unclear under what circumstances an assumption maker, given his or her knowledge of errors in old forecasts, displays assumption drag behaviour when making a new forecast, and when an overreaction occurs. Because a forecaster presumably spends more time analysing recently observed actual trends than error patterns in recent forecasts, assumption drag is more likely to occur than overreaction. Until we know more about the psychology of the forecaster at the time he or she is analysing and extrapolating his or her fertility and mortality data, the only way to prevent assumption drag would be frequent monitoring of observed trends and frequent updating of the forecast. But one should be cautious, and not react immediately to what may turn out to be only short-term changes in trends.

To what extent can the empirical findings in this article be generalized to other countries and other time periods? Many Western countries have experienced a similar fall in birth rates during the 1970s as the countries in Table 1. Similarly, the stagnation in mortality improvement in the 1960s and 1970s is rather widespread. Thus it is quite probable that the forecasts of other industrialized countries produced in the 1960s and 1970s will

show errors in the age structure and in births and deaths of the same magnitude as those reported here. Whether we can generalize to other time periods is more unclear. On the one hand we have seen that actual demographic developments dominate the errors. A further dramatic fall in birth rates, or a return to the high levels of the 1950s and 1960s is not considered likely by population forecasters (Klinger 1991; Szabó 1992). If this expectation is correct, this would suggest forecast errors of decreasing magnitude in the future. On the other hand, Italy and Spain have experienced low fertility levels during the past decade, and these will undoubtly show up as overestimations in the forecasts for the youngest age groups, but not as large as those in Figure 1. To expect a reduction in future errors because of an improvement of forecast quality would be in contradiction with the findings for Norway and the Netherlands, even if these results apply to historical forecasts.

7. References

Adam, A.Y. (1992). The ABS Population Projections: Overview and Evaluation. Journal of the Australian Population Association, 9, 109–130.

Ahlburg, D.A. (1982). How Accurate Are the U.S. Bureau of the Census Projections of Total Live Births? Journal of Forecasting, 1, 365–374.

Alho, J. and Spencer, B. (1991). A Population Forecast as a Database: Implementing the Stochastic Propagation of Error. Journal of Official Statistics, 7, 295–310.

Alho, J. and Spencer, B. (1995). The Practical Specification of the Expected Error of Population Forecasts. Paper presented at the symposium ''Analysis of Errors in Demographic Forecasts with Implications for Policy.'' Koli, Finland, March/April.

Armstrong, J.S. (1985). Long Range Forecasting: From Crystal Ball to Computer (2nd edition). New York: John Wiley.

Ascher, W. (1978). Forecasting: An Appraisal for Policy-makers and Planners. Baltimore and London: The Johns Hopkins University Press.

Bartlema, J. (1987). Developments in Kinship Support Networks for the Aged in the Netherlands: A Social-demographic Analysis. Tilburg, the Netherlands: Catholic University of Brabant (Sociale Zekerheidswetenschap, Rapporten no. 3).

Bretz, M. (1986). Bevölkerungsvorausberechnungen: Statistische Grundlagen und Probleme. Wirtschaft und Statistik, 4, 233–260.

Brunborg, H. (1984). Hvor sikre er befolkningsprognosene? Noen prinsipielle betraktninger om usikkerhet i befolkningsprognoser, Tekniske Rapporter no. 37. Copenhagen: Nordisk Sekretariat.

Cohen, J.E. (1986). Population Forecasts and Confidence Intervals for Sweden: A Comparison of Model-based and Empirical Approaches. Demography, 23, 105–126.

Cruijsen, H. and Keilman, N. (1992). A Comparative Analysis of the Forecasting Process. In National Population Forecasting in Industrialized Countries, N. Keilman and H. Cruijsen (eds.). Amsterdam: Swets and Zeitlinger.

Cruijsen, H. and Zakee, R. (1991). Nationale bevolkingsprognoses in de jaren tachtig: Hoever zaten ze er naast? (''Population Forecasts for the Netherlands During the 1980s: How Far Were They Wrong?'') Maandstatistiek van de Bevolking, 39, 30–39.

De Beer, J. (1988). Predictability of Demographic Variables in the Short Run. European Journal of Population, 4, 283–296.

De Beer, J. (1992). Uncertainty Variants of Population Forecasts. Statistical Journal of the United Nations. ECE, 9, 233–253.

De Jong, A. (1995). Nederlandse bevolkingsprognoses geëvalueerd (''Evaluation of Dutch Population Forecasts''). Maandstatistiek van de Bevolking, 43, 6–9.

Feeney, G. (1990). The Demography of Aging in Japan: 1950–2025. NUPRI Research Paper, Series no. 55. Tokyo: NUPRI.

Field, J. (1990). Past Projections: How Successful? In Population Projections: Trends, Methods and Uses. Occasional Paper 38. London: Office of Population Censuses and Surveys.

George, M.V. and Nault, F. (1991). The Accuracy of Statistics Canada's Demographic Projections. Paper presented at the annual meeting of the PAA, Washington, DC, March.

Hämäläinen, H. (1987). The Accuracy of the Regional Population Projections in Finland in the 1980s. Paper presented at the European Population Conference Jyväskylä, Finland, June.

Hanika, A. (1993). Zur Treffsicherheit von Bevölkerungsvorausschätzungen (Monitoring) (''On the Accuracy of Population Projections (monitoring)''). Statistische Nachrichten, 1, 14–22.

Hansen, H.O. (1993). Elementær demografi (''Elementary Demography''). Copenhagen: Akademisk Forlag.

Keilman, N. (1990). Uncertainty in National Population Forecasting. Amsterdam: Swets and Zeitlinger.

Keilman, N. (1991). Analysing Ex-post Observed Errors in a Series of Population Forecasts. Zeitschrift für Bevölkerungswissenschaft, 17, 411–432.

Keilman, N. and Kucera, T. (1991). The Impact of Forecasting Methodology on the Accuracy of National Population Forecasts: Evidence from The Netherlands and Czechoslovakia. Journal of Forecasting, 10, 371–398.

Keyfitz, N. (1981). The Limits of Population Forecasting. Population and Development Review, 7, 579–593.

Keyfitz, N. (1985). Applied Mathematical Demography (2nd ed.). New York: Springer Verlag.

Klinger, A. (1991). Survey of Recent Fertility Trends and Assumptions Used for Projections. In Future Demographic Trends in Europe and North America: What Can We Assume Today? W. Lutz (ed.). London: Academic Press.

Land, K.C. (1986). Methods for National Population Forecast: A Review. Journal of the American Statistical Association, 81, 888–901.

Lee, R.L. and Tuljapurkar, S. (1994). Stochastic Population Forecasts for the United States: Beyond High, Medium and Low. Journal of the American Statistical Association, 89, 1175–1189.

Long, J.F. (1987). The Accuracy of Population Projection Methods of the U.S. Census Bureau. Paper presented at the Annual Meeting of the PAA, Chicago, April/May.

Long, J.F. (1995). Complexity, Accuracy, and Utility of Official Population Projections. Mathematical Population Studies, 5, 203–216.

Murphy, M. (1987). Population Projections for OECD Countries: A Comparison with Outcomes in the Period 1956–1985. Unpublished paper, London School of Economics and Political Science.

Pollard, J.H. (1966). On the Use of the Direct Matrix Product in Analyzing Certain Stochastic Population Models. Biometrika, 53, 397–415.

Preston, S.H. (1974). An Evaluation of Postwar Mortality Projections in Australia, Canada, Japan, New Zealand, and the United States. WHO Statistical Report, 27, 719–745.

Rideng, A. (1988). The Evaluation of Regional Population Problems. In O. Bertelsen and B. Liebach (eds.) The Eighth Nordic Demographic Symposium, June 1986, Gilleleje, Denmark. Scandinavian Population Studies 8. Copenhagen: The Nordic Demographic Society, 265–279.

Sabatello, E.F. (1988). Population Projections of the Adriatic Populations Toward the 21st Century. In IDRA I, Atti del primo Incontro Demografico delle Regioné Adriatiche (Pescara, novembre 1987). Istituto di Statistica, Università "G. d'Annunzio," Pescara, 283–291.

Schweder, T. (1971). The Precision of Population Projections Studied by Multiple Prediction Methods. Demography, 8, 441–450.

Shaw, C. (1994). Accuracy and Uncertainty of the National Population Projections for the United Kingdom. Population Trends, 77, 24–32.

Smith, S.K. (1987). Tests of Forecast Accuracy and Bias for Country Population Projections. Journal of the American Statistical Association, 82, 991–1003.

Spøhr, H. (1995). Unpublished Figures Regarding the Accuracy of Danish Population Forecasts made Between 1963 and 1994.

Statistiska centralbyrån – SCB (1986). Den framtida befolkningen: Prognos för åren 1986–2025. Demografiska rapporter. Stockholm: Sveriges Officiella Statistik, Statistiska centralbyrån.

Statistiska centralbyrån – SCB (1989). Sveriges framtida befolkning: Prognos för åren 1989–2025. Demografiska rapporter 1989 : 1. Stockholm: Sveriges Officiella Statistik, Statistiska centralbyrån.

Statistiska centralbyrån – SCB (1991). Sveriges framtida befolkning: Prognos för åren 1991–2025. Demografiska rapporter 1991 : 1. Stockholm: Sveriges Officiella Statistik, Statistiska centralbyrån.

Statistics Canada (1994). Population Projections for Canada, Provinces and Territories 1993–2016. Ottawa: Statistics Canada (by M.V. George, M.J. Norris, F. Nault, S. Loh and S.Y. Dai).

Statistics Norway (1994). Framskriving av folkemengden 1993–2050: Nasjonale og regionale tall ("Population Projections 1993–2050: National and Regional Figures"). NOS C 176. Oslo: Statistics Norway.

Stoto, M.A. (1983). The Accuracy of Population Projections. Journal of the American Statistical Association, 78, 13–20.

Szabó, K. (1992). International Comparison of Fertility Assumptions. In National Population Forecasting in Industrialized Countries, N. Keilman and H. Cruijsen (eds.). Amsterdam: Swets and Zeitlinger.

Texmon, I. (1992). Norske Befokningsframskrivinger 1969–1990 ("Norwegian population projections 1969–1990). In Mennesker og modeller: Livsløp og kryssløp, O. Ljones, B. Moen and L. Østby (eds.). Sosiale og Økonomiske Studier no. 78. Oslo: Statistics Norway.

Texmon, I. and Keilman, N. (1990). An Evaluation of Norwegian Population Forecasts:

Experiences from the Period 1969–1987. In Nordic Seminar on Forecasting, Drammen, 24–26 April. Tekniske Rapporter 53. Copenhagen: Nordisk Statistisk Sekretariat.

U.S. Bureau of the Census – USBC (1984). Projections of the Population of the United States by Age, Sex, and Race: 1983–2080. Current Population Reports, Series P-25, no. 952.

U.S. Bureau of the Census – USBC (1989). Projections of the Population of the U.S. by Age, Sex, and Race: 1988–2080. Current Population Reports, Series P-25, no. 1018 (by G. Spencer).

U.S. Bureau of the Census – USBC (1992). Population Projections of the United States, by Age, Sex, Race, and Hispanic Origin: 1992 to 2050. Current Population Reports, Series P-25, no. 1092 (by J.C. Day).

Vallin, J. (1989). L'avenir de l'espérance de vie vu à travers les projections de l'INSEE. Population 44, 930–936.

Van Poppel, F. and de Beer, J. (1993). Evaluation of Standard Mortality Projections for the Elderly. Paper presented at IUSSP Seminar on Health and Mortality Trends Among Elderly Populations: Determinants and Implications. Sundai, Japan, June.

Received July 1995
Revised February 1996