Discussion Paper

# Quasi stochastic population forecasts



2017 | 01

Coen van Duin

# Content

## Summary

Stochastic techniques for estimating uncertainty intervals for demographic forecasts are still not widely used by statistical institutes. One reason for this may be the complexity of the Monte Carlo approach, which requires the calculation of a thousand or more variants of the deterministic forecast. This paper discusses a technique to estimate forecast intervals for demographic indicators from only 6 variants. Good agreement is found between the forecast intervals obtained with this technique and from a stochastic forecast. This method does not provide a full probability distribution for the forecasts, however. Also, it cannot be used to compute intervals for demographic flows, although intervals for time-cumulated flows can be obtained.

## Keywords

Population forecasts, stochastic forecasting, demography

# 1. Introduction

It has long been realized that deterministic population forecasts are of limited use without an indication of their uncertainty. A deterministic point forecast only gives the most probable value of, for instance, the population size in 50 years' time, but no information on how close to this value the actual population size is likely to be. Already in the early 1970s, Keyfitz argued that population forecasts should preferably be made in the form of a probability function (Keyfitz, 1972). Alho and Spencer (1991) attribute the earliest population forecasts with a probabilistic interpretation to the Finnish Central Statistical Office (Hyppölä et al., 1949). Since then, different techniques have been developed to produce such stochastic forecasts. Alho and spencer (1985) developed the analytic propagation of error approach. Pflaumer (1986) used Monte Carlo simulation. In the 1990s, stochastic population forecasts were published for a number of countries: Austria (Hanika et al., 1997; Lutz and Scherbov, 1998b), Finland (Alho, 1998), Germany (Lutz and Scherbov, 1998a), the Netherlands (Alders and de Beer, 1998), and the United States (Lee and Tuljapurkar, 1994).

Despite this long history, stochastic techniques are still not widely used in official population forecasts. Most statistical agencies produce only deterministic forecasts. An exception is Statistics Netherlands, which has been publishing stochastic forecasts since 1998 (De Beer and Alders, 1999). More recently, the United Nations has started publishing probabilistic forecasts (UN, 2014). Their 2012 en 2015 forecasts take into account uncertainty in fertility and mortality, but not in international migration.

Many statistical offices do compute a number of projection variants in addition to the official forecast to present alternative futures. Typically, these explore what will happen if fertility turns out higher or lower than assumed in the forecast, or mortality, or net migration. In the population projections of the German Federal Statistical Office (Pötzsch and Rößger, 2015), all combinations of high and low assumptions for these 3 components are explored –without publishing a forecast, or median variant.

While such variants give insight into the sensitivity of the forecast to the input assumptions and make users aware of the fact that forecasts are uncertain, they do not provide a consistent estimate of the forecast uncertainty. Firstly, the range between the results of the high and low variant typically covers a smaller probability in the short term than in the long term (Lee and Tuljapurkar, 1994). Secondly, the results of the different projection variants are not combined to give one high and one low margin that combines the effects of uncertainty in each of the components.

One reason the official population forecasts are still largely deterministic may be the perceived complexity of the stochastic approach. This paper investigates if it is possible to reproduce the main results of stochastic forecasts, forecast intervals for stock variables such as number of residents or grey pressure, using the familiar variant approach. This is done by adding two elements. First: a method to construct

time paths for the input assumptions of high and low fertility-, mortality- or migration variants in such a way that their output range for stock variables is similar to the confidence interval obtained from a stochastic forecast (in which only the uncertainty in the particular component is taken into account). Second: a method to combine the results from the 6 fertility, mortality and migration variants into a single, quasi stochastic, uncertainty interval.

In section 2, a simplified version of the stochastic forecast method used at statistics Netherlands is discussed. Section 3 introduces the quasi stochastic method. In section 4, the quasi stochastic forecast intervals for the 2012-based forecasts of Statistics Netherlands are compared to stochastic intervals computed using the method described in section 2. These results are discussed (section 5) and some conclusions are drawn (section 6).

# 2. Stochastic method

In this paper, the 2012-based stochastic population forecast of Statistics Netherlands (Van Duin and Stoeldraijer, 2013) are used as a benchmark to which the results of the quasi stochastic method are compared. This section discusses a simplified version of this stochastic forecast. This simplified version does not distinguish the residents by their country of birth. It uses number of emigrants as input, instead of county of birth-dependent emigration rates. Instead of varying both the assumptions for immi- and emigration, only those for net migration are varied. In section 4, the output of the quasi stochastic method is compared to the results of this simplified stochastic model. For both methods, the same assumptions are used for the time series models that generate the input errors in the vital rates and the migration numbers.

The population forecast is computed using a standard cohort-component model. Starting point is the national population by age and sex at the end of the last observed year ($t_0$=2011). The forecast is calculated by running a macro-simulation on this starting population. Each year, immigrants are added to and emigrants removed from the population, births and deaths are generated based on the fertility and mortality rates and all members of the population are aged by one year. In this way, populations for subsequent years are generated.

The cohort-component model is a bookkeeping model. Given the correct input, it will produce correct output. Forecast uncertainty arises because the correct values for the input parameters are not exactly known. In many cases, the starting population is known accurately, from a population register or a census, and the main source of uncertainty is the fact that the future development of fertility, mortality and migration cannot be predicted.

In a deterministic forecast, the most likely time paths are specified for the future development of the fertility and mortality rates and the numbers of migrants. In the stochastic approach, in addition to this, time series models are specified for the input errors in one indicator for each component. For the component fertility, the total fertility rate (*TFR*) is used as indicator, for migration, the net migration number (*N*), for mortality, life expectancy at birth ($e_0$).

In the stochastic forecast of Statistics Netherlands, the input errors in the total fertility rate and life expectancy at birth are modelled as a random walk.

$$\Delta TFR\,(t) = \Delta TFR\,(t-1) + \varepsilon_{TFR}\,(t), \qquad (1)$$

$$\Delta e_0(g,t) = \Delta e_0(g,t-1) + \varepsilon_{e_0}(t), \qquad (2)$$

where $\Delta TFR(t) = TFR(t) - \hat{TFR}(t)$ is the difference between the realised and assumed value of the *TFR* (the hat is used here to indicate the value in the deterministic forecast). $\Delta e_0(t)$ is the value of the input error in life expectancy for gender $g$. The ε's in (1) and (2) are error terms drawn from a symmetrical normal

distribution with time-independent standard deviations $\sigma_{TFR}$ and $\sigma_{e_0}$. The error terms are uncorrelated in time and uncorrelated between components. To ensure that the life expectancies for men and women do not cross in any variant, a gender-independent error term is used in (2).

The time series models (1) and (2) imply that the width of the 67% intervals for $\Delta TFR$ and $\Delta e_0$increase as the square root of forecast lead time.

$$\begin{aligned} \mathrm{P}_{5/6}[\Delta TFR\,(t_0+k)] - \mathrm{P}_{1/6}[\Delta TFR\,(t_0+k)] &= 0.967\cdot 2\sigma_{TFR}\sqrt{k}, \\ \mathrm{P}_{5/6}[\Delta e_0(g,t_0+k)] - \mathrm{P}_{1/6}[\Delta e_0(g,t_0+k)] &= 0.967\cdot 2\sigma_{e_0}\sqrt{k}, \end{aligned} \tag{3}$$

where $\mathrm{P}_x[Y]$ denotes the percentile of rank 100*$x$ of the distribution of $Y$.

The input errors in the net migration number are assumed to be generated by an AR(1) time series model

$$\Delta N(t) = \varphi \Delta N(t-1) + \varepsilon_N(t). \tag{4}$$

The autocorrelation parameter φ describes to what extent input errors persists into the next year. If φ equals zero, input errors do not persist and (4) describes white noise with no autocorrelation. If φ equals 1, (4) describes a random walk. The input errors for net migration in the 2012-based stochastic forecast of Statistics Netherlands can be modelled[1] as an AR(1) time series with φ =0.87. Based on (4) the interval width of $\Delta N$ increases with time as

$$\begin{aligned} \mathrm{P}_{5/6}[\Delta N(t_0+k)] - \mathrm{P}_{1/6}[\Delta N(t_0+k)] &= 0.967\cdot 2\sigma_N\sqrt{\sum_{l=0}^{k-1}\varphi^{2l}} \\ &= 0.967\cdot 2\sigma_N\sqrt{\frac{1-\varphi^{2k}}{1-\varphi^2}}. \end{aligned} \tag{5}$$

For $0 < \varphi < 1$, this means that the magnitude of the input errors in net migration increase more rapidly in the short term and more slowly in the long term than those for fertility and mortality.

There are three main methods for estimating the uncertainty in the vital rates and migration number: time-series extrapolation, expert judgement and extrapolation of historical errors in the forecast assumptions (Keillman, 2002; Alders et al., 2005). These methods can also be combined. For instance: expert judgement is used to determine the time period on which the time series model for the vital rates is estimated, or patterns in past input errors are used to check the validity of the

[1] In the stochastic model of Statistcs Netherlands, the input errors in the number of immigrants (by country of birth) and in the emigration rates (by country of birth) are modelled separately as AR(1) time series with a non-normal distribution for the error term. Nevertheless, the distribution of the net number of migrants obtained from the output of this model is approximately normal and symmetric.

assumed time series for the input errors. Table 1 shows the assumptions used in the 2012 stochastic forecast of Statistics Netherlands.

Once the time series models for the errors in the input components are specified, the next step is to use the Monte Carlo method. A large number ($N_{var}$) of projection variants is created by

1. Drawing values of $\varepsilon_{TFR}(i,t)$, $\varepsilon_{e_0}(i,t)$ and $\varepsilon_N(i,t)$ at random from their respective distributions. The index i indicates the variant. $N_{var}$=10,000 was used for the results shown in section 4 .
2. Calculating time series $\Delta\tilde{TFR}(i,t)$, $\Delta\tilde{e}_0(i,t)$ and $\Delta\tilde{N}(i,t)$ for the input errors in each variant using these random numbers (the tilde is used here to indicate the value in a projection variant).
3. Creating input values for the fertility rates, mortality risks and migration numbers for each variant that reproduce these input errors. To reproduce the errors $\Delta\tilde{TFR}(i,t)$ and $\Delta\tilde{e}_0(i,t)$, the fertility rates and mortality risks from the deterministic forecast are scaled with an age-independent factor. The input errors for net migration $\Delta\tilde{N}(i,t)$ are reproduced by adjusting the number of immigrants from the deterministic forecast with an age-independent factor.[2]
4. Running the cohort-component model $N_{var}$ times using the input generated in this way and storing the output.

Instead of one value for each output variable (e.g. population size or grey pressure), this procedure yields $N_{var}$ values. The distribution of these output values is interpreted as the probability distribution for the future values of the output variable. By construction, the median of this distribution corresponds to the value according to the deterministic forecast. Forecast intervals are obtained from the percentiles of the distribution. For instance, the upper and lower margins for the 67% interval are estimated by

$$\begin{aligned} \mathrm{UM}_{67\%}[Y] &= \mathrm{P}_{1/6}[\{\tilde{Y}_i\}, , \\ \mathrm{LM}_{67\%}[Y] &= \mathrm{P}_{5/6}[\{\tilde{Y}_i\}], \end{aligned} \qquad (6)$$

where $\mathrm{P}_x[\{Y_i\}]$is estimated from the distribution of values of the output variable *Y* accros the $N_{var}$ projection variants.

**Table 1: assumptions for the uncertainty in fertility, life expectancy and net migration.**

| | $TFR$ | $e_0$ | $N$ |
|---|---|---|---|
| | **children/woman** | **years** | **1,000 migrants** |
| $\sigma$ | 0.04 | 0.4 | 15.3 |
| width (67%, 50 years) | 0.53 | 5.4 | 59.8 |

[2] Alternative, some mixture of an adjustment of the number of immigrants and emigrants could be used. However, since the age and gender structure of immigrants and emigrants is quite similar for The Netherlands, this does not strongly influence the results.

# 3. Quasi stochastic method

Is it possible to approximate the forecast intervals obtained from the stochastic method using only 6 projection variants: one high and one low variant for each component? To answer this question, it is useful to first consider the simpler problem of approximating with two variants the intervals from a stochastic forecast in which only one component is varied.

## 3.1 Single component uncertainty

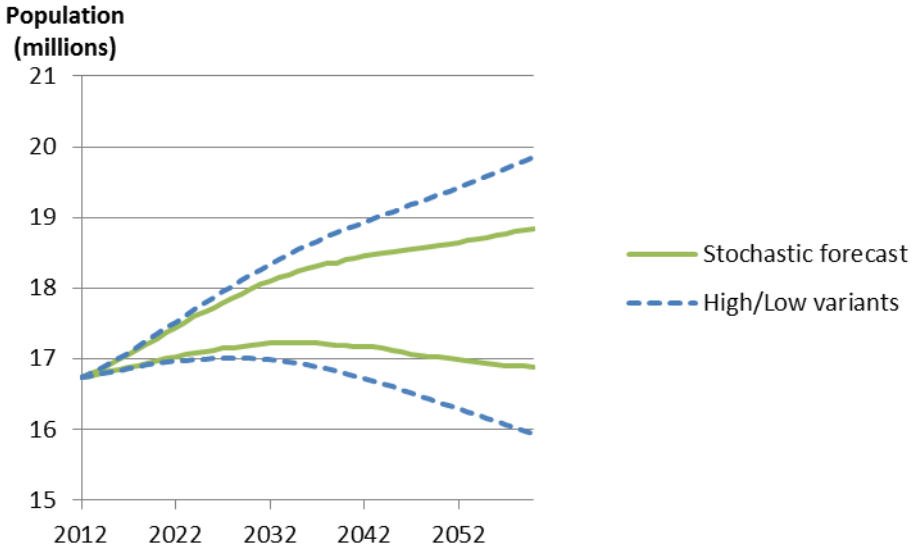
Stochastic forecasts can also be used to estimate the contribution of a single component to the uncertainty of the results. In that case, the input in the variants is adjusted only for the component under consideration. So all variants would have, for instance, the same input rates for mortality and fertility, but different numbers of migrants. Is it possible to approximate the projection intervals calculated in this way using only one high and one low variant?

As a first try, we could use the 67% upper and lower margins for the indicator of the component that we want to study. Suppose we do this for net migration. The two variants, MigHigh and MigLow, would have input based on the following indicators.

$$\begin{aligned}
&\tilde{TFR}(\text{MigHigh}\quad,t) = \tilde{TFR}(\text{MigLow}\quad,t) = \hat{TFR}(t),\\
&\tilde{e}_0(\text{MigHigh}\quad,g,t) = \tilde{e}_0(\text{MigLow}\quad,g,t) = \hat{e}_0(g,t),\\
&\tilde{N}(\text{MigHigh}\quad,t) = \hat{N}(t) + \text{P}_{5/6}[\Delta N(t)],\\
&\tilde{N}(\text{MigLow}\quad,t) = \hat{N}(t) + \text{P}_{1/6}[\Delta N(t)],
\end{aligned} \tag{7}$$

Figure 1 shows the intervals for population size for the Netherlands, for the 2012-based forecast, obtained from these two variants and from the stochastic forecast in which only net migration is varied. The population size according to the MigHigh variant is consistently higher than the 67% upper margin according to the stochastic forecast, the MigLow variant consistently lower than the 67% lower margin. The only exception is for the population one year into the forecast, where they are nearly identical. At the forecast horizon in 2060, the width of the interval from the variants is too high by a factor 2.

**Figure 1: Migration-induced uncertainty in the population size, from a stochastic forecast and from High/Low-variants without narrowing.**




The intervals obtained from (7) are too wide because of short-term fluctuations in the net migration that average out after a number of years. Such fluctuations contribute to the width of the interval for net migration (5), but not, or hardly, to the width of the interval for population size. As an example, suppose that net migration is 5,000 lower than expected in 2013 and 5,000 higher than expected in 2014. The input error in *N* in 2014 is 5,000, but the induced forecast error for population size at the end of 2014 is nearly zero. The forecast error in population size in year t does not depend on the input error in net migration in year t, but on the cumulated input error in net migration over the forecast years up to t. The same is true for the induced forecast error in the number of births due to input errors in net migration.

To correct for the effect of short-term fluctuations in the input errors, a time-depend factor is added to (7) to narrow the interval between the MigHigh- and MigLow variant

$$\begin{aligned}\tilde{N}(\text{MigHigh}\quad,t) &= \hat{N}(t) + \psi_{\text{high}}(t)\,\text{P}_{5/6}[\Delta N(t)],\\ \tilde{N}(\text{MigLow}\quad,t) &= \hat{N}(t) + \psi_{\text{low}}(t)\,\text{P}_{1/6}[\Delta N(t)],\end{aligned} \tag{8}$$

Because the forecast error in population size depends on the cumulated input error in net migration over the preceding years, it is imposed that the cumulated input error for the two variants corresponds to the upper and lower margin of the 67% interval for the cumulated input error according to the time series. So, for the high variant in forecast year *t*:

$$\sum_{k=1}^{t-t_0}\Delta\tilde{N}(\text{MigHigh}\quad,t_0+k) = \text{P}_{5/6}\left[\sum_{k=1}^{t-t_0}\Delta N(t_0+k)\right] \tag{9}$$

Inserting the form (8) into (9) yields

$$\sum_{k=1}^{t-t_0} \psi_{\text{high}}(t_0+k)\,\mathrm{P}_{5/6}\left[\Delta N(t_0+k)\right] = \mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0} \Delta N(t_0+k)\right]. \tag{10}$$

By splitting of the last term in the summation on the left hand side, an expression for the narrowing factor can be obtained

$$\begin{aligned}
&\psi_{\text{high}}(t)\,\mathrm{P}_{5/6}\left[\Delta N(t)\right] + \sum_{k=1}^{t-t_0-1} \psi_{\text{high}}(t_0+k)\,\mathrm{P}_{5/6}\left[\Delta N(t_0+k)\right] \\
&= \mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0} \Delta N(t_0+k)\right] \\
&\Leftrightarrow \psi_{\text{high}}(t)\,\mathrm{P}_{5/6}\left[\Delta N(t)\right] + \mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0-1} \Delta N(t_0+k)\right] = \mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0} \Delta N(t_0+k)\right] \\
&\Leftrightarrow \psi_{\text{high}}(t) = \frac{\mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0} \Delta N(t_0+k)\right] - \mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0-1} \Delta N(t_0+k)\right]}{\mathrm{P}_{5/6}\left[\Delta N(t)\right]}.
\end{aligned} \tag{11}$$

For $\psi_{low}(t)$, the same expression is obtained with $\mathrm{P}_{5/6}[\dots]$ replaced by $\mathrm{P}_{1/6}[\dots]$. The expression (11) does not depend on which time series model is assumed for the input errors. For an AR(1) model (eq.(4)) with a symmetric distribution of the error term, the narrowing factor derived from (11) is given by (see the appendix for a derivation)

$$\psi_{\text{high}}(t) = \psi_{\text{low}}(t) = \frac{\sqrt{\sum_{j=1}^{t-t_0}\left(\sum_{k=j}^{t-t_0} \varphi^{k-j}\right)^2} - \sqrt{\sum_{j=1}^{t-t_0-1}\left(\sum_{k=j}^{t-t_0-1} \varphi^{k-j}\right)^2}}{\sqrt{\sum_{k=0}^{t-t_0-1} \varphi^{2k}}}.. \tag{12}$$

This expression can easily be calculated numerically. Figure 2 shows the narrowing factor for different values of the autocorrelation parameter.

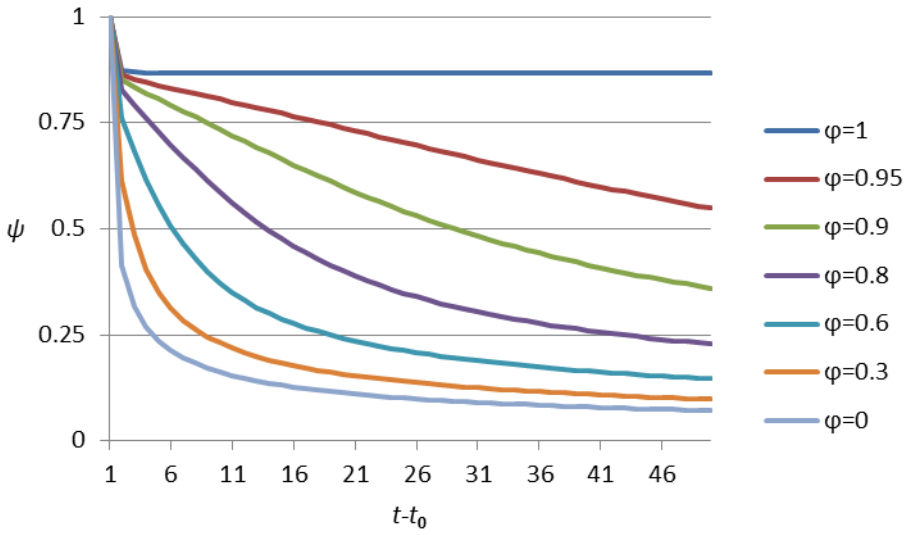
If the input errors have no autocorrelation ($\varphi$=0), only the terms with $\varphi^0$ contribute to the summations in (12) and the narrowing factor is given by

$$\psi_{\varphi=0}(t) = \sqrt{t-t_0} - \sqrt{t-t_0-1}\,. \tag{13}$$

For $\varphi$=1, strong autocorrelation, the AR(1) series becomes a random walk and the narrowing factor becomes nearly constant after the first forecast year.

$$\psi_{\varphi=1}(t) = \begin{cases} 1; t = t_0+1 \\ \approx 0.87; t > t_0+1 \end{cases} \tag{14}$$

**Figure 2: Narrowing factor for an AR(1) time series model for different values of the autocorrelation parameter.**




To derive variants that describe the contribution of fertility to the uncertainty in the projections, the same procedure is followed. The following assumption is used for the *TFR* in the FertHigh and the FertLow variant

$$\begin{aligned} \widetilde{TFR}(\text{FertHigh} \quad ,t) &= \hat{TFR}(t) + \psi_{\varphi=1}(t)\,\mathrm{P}_{5/6}[\Delta TFR\ (t)], \\ \widetilde{TFR}(\text{FertLow} \quad ,t) &= \hat{TFR}(t) + \psi_{\varphi=1}(t)\,\mathrm{P}_{1/6}[\Delta TFR\ (t)]. \end{aligned} \tag{15}$$

where the narrowing factor with $\varphi$=1 (eq. 14) is used because the input error for the *TFR* is modelled as a random walk (eq. (1)).

The impact of input errors in the *TFR* on the projected numbers of births depends on the size of the population at risk (women in the fertile ages) which varies from year to year. This is different from the case of input errors in net migration. However, as long as the year-to-year fluctuations in the population at risk are much smaller, in relative terms, than the input errors in the *TFR*, it should still be a good approximation to use the $\psi$ derived from (11).

For mortality, the situation is somewhat more complex, because the indicator (life expectancy) does not have a linear relation with the number of deaths. The high and low mortality variant without narrowing use the following assumptions (analogous to (15)):

$$\begin{aligned} \tilde{e}_0(\text{MortHigh} \quad ,t) &= \hat{e}_0(t) + \mathrm{P}_{1/6}[\Delta e_0(t)]], \\ \tilde{e}_0(\text{MortLow} \quad ,t) &= \hat{e}_0(t) + \mathrm{P}_{5/6}[\Delta e_0(t)]. \end{aligned} \tag{16}$$

For both variants, an adjustment factor $F(g,t)$ is calculated that, when applied to mortality risks $\tilde{q}(g,x,t)$ ($g$=gender, $x$=age) reproduces the value for life expectancy

in that variant according to (16). The narrowing factors are applied to the values of $F(g,t)$ in the MortLow and MortHigh variant

$$\begin{aligned}\widetilde{F}(\text{MortHigh}\quad,g,t) &= 1+\psi_{\varphi=1}(t)\,\mathrm{P}_{5/6}[\Delta F(g,t)],\\ \widetilde{F}(\text{MortLow}\quad,g,t) &= 1+\psi_{\varphi=1}(t)P_{1/6}[\Delta F(g,t)],\end{aligned} \tag{17}$$

where $\Delta F(g,t) = F(g,t) - 1$, because *F*=1 in the case that there is no input error in life expectancy. $\mathrm{P}_{5/6}[\Delta F(g,t)]$ is the value of $\Delta F(g,t)$ that produces $\mathrm{P}_{1/6}[\Delta e_0(g,t)]$ $\Delta F(g,t)$when applied to the mortality risks and $\mathrm{P}_{1/6}[\Delta F(g,t)]$ is the value that produces $\mathrm{P}_{5/6}[\Delta e_0(g,t)]$. This follows from the fact that there is a one to one correspondence between the adjustment factor and $e_0$ (higher adjustment factor, lower life expectancy). The narrowing factor with $\varphi$=1 is used because, if $\Delta e_0(g,t)$ follows a random walk, the same is approximately true for $\Delta F(g,t)$, for sufficiently small deviations.

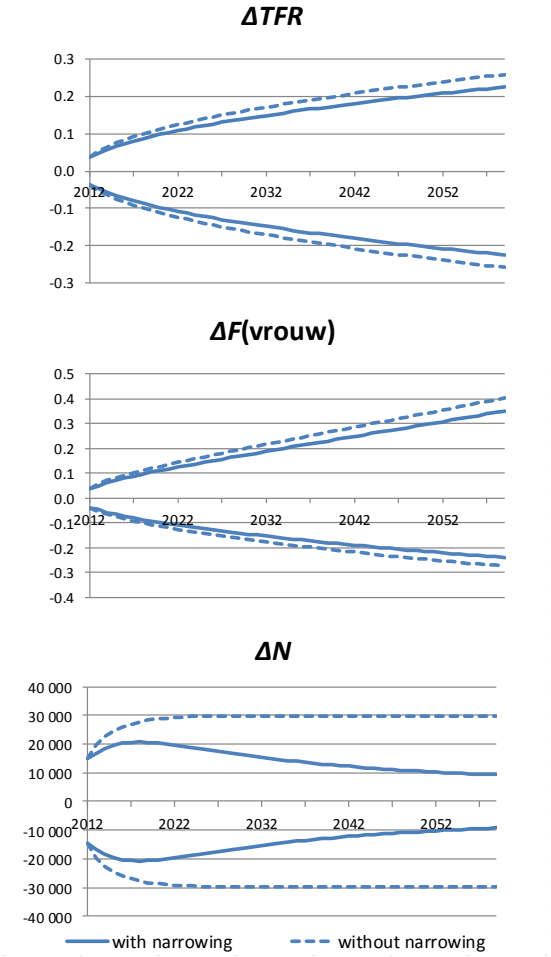
As is the case for fertility and births, the impact of the input errors in the mortality rates on the number of deaths depends on the population at risk, which is not a constant. Again, it is expected that this will cause no large error in the estimation of the narrowing factor, because the relative year-to-year fluctuations in F are much larger than those in the population at risk. For mortality there is a second concern, which is that the distribution of $\Delta F(g,t)$is not normal (it is approximately lognormal). This means that the expression (12) does not necessarily give a good estimate for the narrowing factors. To test this, the narrowing factors were estimated from the 10,000 input series $\tilde{F}(i,g,t)$ of the stochastic population forecast. Equation (11) was used to estimate $\psi(t)$, since this expression also holds if the distribution is not normal. The estimates of $\psi(t)$ obtained in this way were quite noisy, but showed no significant time dependence after the first forecast year. Assuming a constant underlying value of $\psi(t) = c$ for $t > t_0 + 1$ yielded the estimates $c_{low} = 0.86 \pm 0.01$ and $c_{high} = 0.88 \pm 0.02$ (the margins indicate the 95% confidence interval). This suggests that the estimate $\psi(t > t_0 + 1) = 0.87$ from equation (12) is still a good approximation.

Finally, with mortality there is an effect that input errors which lead to more deaths in one year lead to fewer deaths in the next year (because people cannot die twice) and vice versa (because death is only postponed, not avoided). This second order dampening effect is not taken into account in these calculations. The time paths for the input errors in the Mort-variants have nearly perfect autocorrelation, while the paths in the stochastic forecasts are more erratic. As a result, the dampening effect will be stronger for the 2 Mort-variants than for the 10,000 variants of the stochastic forecast. This will tend to make the intervals obtained from the Mort-variants too narrow. The difference could be small, however, because a time series with strong autocorrelation (random walk) is used to generate the input errors in the mortality rates for the stochastic forecast.

The resulting high and low trajectories for $\Delta TFR(\mathrm{t})$, $\Delta F(\mathrm{Female},t)$ and $\Delta N(\mathrm{t})$ are shown in figure 3, with and without narrowing. Notice that the interval between the high and low variant of *N* decreases with forecast duration in the later years of the forecast. This is very different from the high/low-variants that are usually used, which

aim to describe the uncertainty in the input component $N(t)$. The variants used here aim to describe the uncertainty in the time-cumulated input component $\sum_{k=1}^{t-t_0} \Delta N(t_0 + k)$.

**Figure 3: assumptions used in the variants for fertility, mortality and net migration.**




## 3.2 Multi component uncertainty

In the stochastic method, a distribution of output values is generated by independently drawing input errors for fertility, mortality and migration and computing a variant for each combination of input errors. Something similar is done in the quasi stochastic method. A much simpler distribution of input errors is used, consisting of only one high and one low input error for each component, both with probability ½. Since there are 3 components, there are only $2^3 = 8$ possible combinations of input errors, so the distribution of input errors only has 8 values, each with probability 1/8. Upper and lower margins are derived from the first and second moment of this distribution of input errors

$$\Delta\tilde{\tilde{Y}}_{\mathrm{UM}}(t) = \mathrm{M}[\Delta Y(t)] + \sqrt{\mathrm{M}[\Delta Y(t)^2] - \mathrm{M}[\Delta Y(t)]^2},$$
$$\Delta\tilde{\tilde{Y}}_{\mathrm{LM}}(t) = \mathrm{M}[\Delta Y(t)] - \sqrt{\mathrm{M}[\Delta Y(t)^2] - \mathrm{M}[\Delta Y(t)]^2}, \quad (18)$$

with

$$\mathrm{M}[\Delta Y(t)^a] = \frac{1}{8}\sum_{r_1=\mathrm{Low}}^{\mathrm{High}}\sum_{r_2=\mathrm{Low}}^{\mathrm{High}}\sum_{r_3=\mathrm{Low}}^{\mathrm{High}} \Delta\tilde{Y}(r_1, r_2, r_3, t)^a, a = 1,2. \quad (19)$$

where $\Delta\tilde{Y}(r_1, r_2, r_3, t)$ is the forecast error in output variable *Y* for a variant in which fertility has an input error with direction $r_1$ (high or low), mortality with direction $r_2$ and migration with direction $r_3$.

For the stochastic forecast, the 67% margins derived from (18) and from the percentiles of the output distribution (6) are very similar.

As long as the effect of the input errors in each component on the population at risk for the other components is small compared to the size of that population, it is a good approximation to assume that there is no interaction between input errors in different components. This means that the forecast error generated by a combination of input errors in the three components is the same as the sum of the forecast errors generated each input error separately. So

$$\Delta\tilde{\tilde{Y}}(r_1, r_2, r_3, t) \approx \sum_{s=1}^{3} \Delta\tilde{\tilde{Y}}_s(r_s, t), \quad (20)$$

where $\Delta\tilde{Y}_S(r_S, t)$ is the forecast error in a variant in which only component *s* has an input error (with sign $r_S$). These are the variants discussed in the previous section. Using this approximation, the number of variants that has to be computed is further reduced from $2^3 = 8$ to $2 \cdot 3 = 6$. The first and second moment of the distribution of forecast errors are estimated by

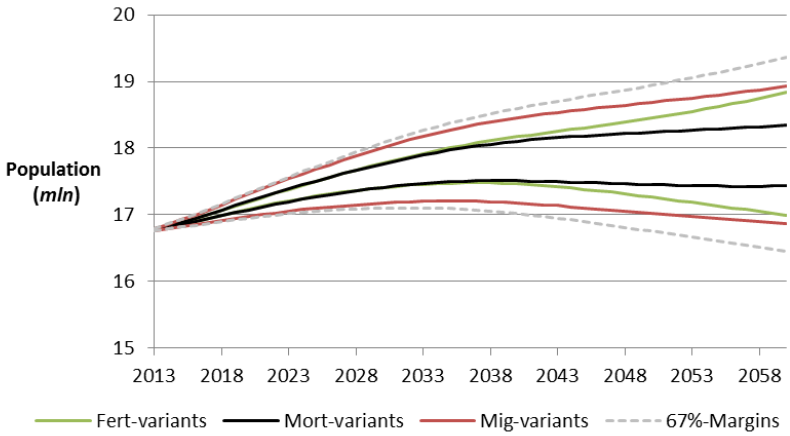
$$\mathrm{M}[\Delta Y(t)^a] \approx \frac{1}{8}\sum_{r_1=\mathrm{Low}}^{\mathrm{High}}\sum_{r_2=\mathrm{Low}}^{\mathrm{High}}\sum_{r_3=\mathrm{Low}}^{\mathrm{High}} \left(\sum_{s=1}^{3} \Delta\tilde{\tilde{Y}}_s(r_s, t)\right)^a, a = 1,2. \quad (21)$$

Figure 4 shows the population size for the Netherlands according to the 6 variants and the resulting 67% forecast interval obtained from (21). For the short term, the forecast uncertainty for population size is mainly due to uncertainty in migration. The upper and lower margin of the forecast interval is therefore close to the value from the MigHigh- and MigLow variant. Further into the forecast, the contribution of fertility to the forecast uncertainty becomes more important and the forecast interval becomes wider than the interval between the Mig-variants.

For different output variables, the contribution of the different components to forecast uncertainty is different. For the population aged 0-19, fertility and migration contribute about equally to the uncertainty in the short term, but fertility dominates in the long term. For the working age population, migration uncertainty dominates

throughout the forecast period. For the old age population, mortality is the main source of uncertainty.

**Figure 4: Population size, output from the 6 variants and pseudo stochastic forecast interval (67%).**



# 4. Results

The quasi stochastic method is applied to the 2012-based population forecasts for Statistics Netherlands and the resulting uncertainty intervals are compared to those for the simplified stochastic forecast based on the same time series for the input errors. For this stochastic forecast, 10 thousand variants were calculated using Monte Carlo simulation. Figure 5 shows results for six demographic indicators: the total population, population by large ages groups (0-19 years, 20-64 years and 65 years and older), grey pressure and green pressure. In addition to the 67% forecast intervals, 95% intervals are shown. The quasi stochastic 95% intervals were obtained from six additional variants based on the 2.5% and 97.5% percentiles of the distribution of input errors.

The general impression from these figures is that good agreement is found between the intervals obtained from the quasi stochastic and stochastic method. In absolute terms, the difference in the width of the forecast intervals for both methods increases with forecast horizon and is larger for 95% intervals than for 67% intervals.

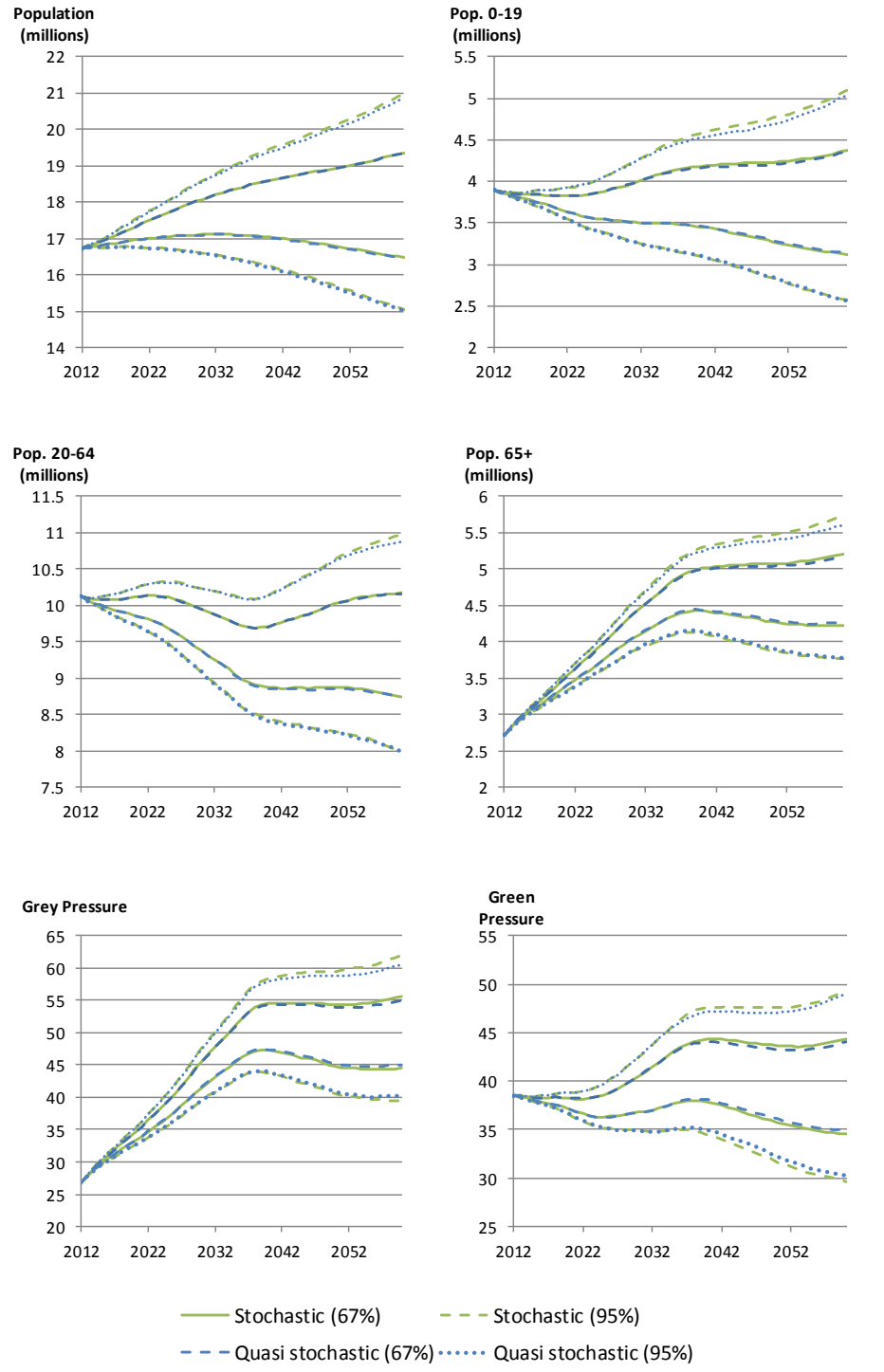
To give a more quantitative assessment of the similarity between the intervals from both methods, we look at the following summary measure for the relative difference between the width of the intervals in the forecast period $t_0 + i$ to $t_0 + j$

$$d\,(i-\ j\ \text{years}\ ) = \frac{\sum_{t=t_0+i}^{t=t_0+j} W_{QS}\,(t)}{\sum_{t=t_0+i}^{t=t_0+j} W_S\,(t)} - 1, \qquad (22)$$

where $W_S(t)$ is the width in forecast year *t* of the stochastic intervals and $W_{QS}(t)$ of the quasi stochastic intervals. By using sums of widths in the numerator and denominator of (22), the forecast years near the end of the selected period, which have the broadest intervals, are given the most weight in *d*. To distinguish between intervals for short and long-term forecasts, we look at the interval 0-25 years and 26-48 years (table 2).

For short term forecasts, up to 25 years, the relative difference in width between both types of intervals is at most 4% for the 6 indicators considered here. For long term forecasts, the discrepancy increases. Still, the maximum discrepancy found is only 8% of the interval width. The method seems to work equally well for 67% and 95% intervals. The forecast intervals for green and grey pressure are reproduced less accurately, for long term forecasts, than those for the population at young and old ages. This suggests that the quasi stochastic method is better at estimating uncertainty intervals for population counts than for ratios of population counts of different age groups. The quasi stochastic intervals are somewhat too narrow for the younger and older ages.

**Figure 5: Quasi-stochastic and stochastic forecast intervals for 6 demographic indicators.**



The intervals for the total population are reproduced more accurately, on the whole, than those for age groups. What happens if the method is used for much smaller age groups? For 5-year age groups, the quasi stochastic method still reproduces quite well the general age-structure of the uncertainty intervals (figure 6). On average, the quasi stochastic intervals per age group in 2060 are 6% narrower than the stochastic ones, but there are substantial differences between the age groups. For the ages 45-54, the stochastic intervals are 16% narrower, for the ages 65-74, 8% broader. The quasi stochastic method tends to overestimate the uncertainty of the birth cohorts which are in their early twenties in the first years of the forecast, the ages at which immigration numbers are highest. For the other cohorts, uncertainty is underestimated.

The quasi stochastic method cannot be used to estimate margins for demographic flows ( births, deaths, net migration). The margins obtained for the flows are too narrow, especially for deaths and net migration (of course, the margins for net migration are known already from the input assumptions eq.(5)). See table 3. The method does, however, yield good estimates for the margins for the cumulated flows up to the forecast year: the sum of all births, deaths, or migrants from the starting year of the forecast up to year *t*.

**Figure 6: Forecast intervals for population by 5 year age group in 2060**

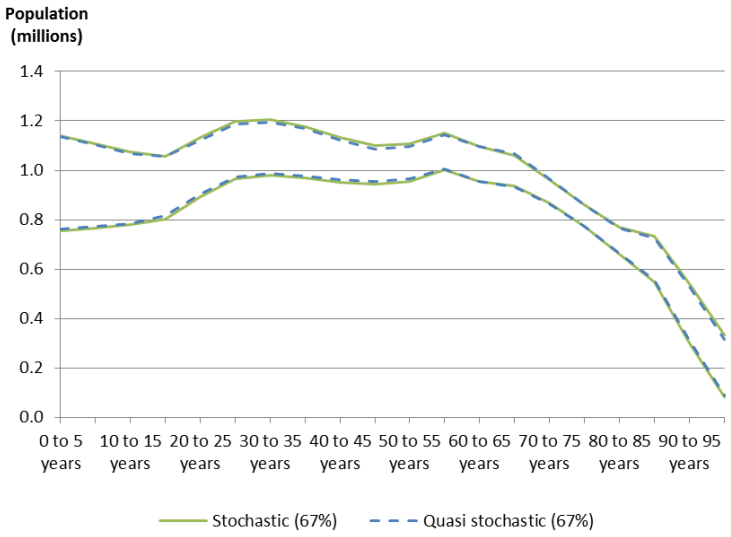


**Table 2: Relative difference between the widths of the quasi stochastic and stochastic forecast intervals for population stocks.**

| | Pop. | Pop. 0-19 | Pop. 20-64 | Pop. 65+ | Grey press. | Green press. |
|---|---|---|---|---|---|---|
| | % | | | | | |
| 67% interval | | | | | | |
| *d*(0-25 years) | 1 | 0 | -1 | -4 | -4 | -1 |
| *d*(25-48 years) | 1 | -4 | 1 | -6 | -8 | -8 |
| 95% interval | | | | | | |
| *d*(0-25 years) | -1 | -1 | -1 | -4 | -3 | -1 |
| *d*(25-48 years) | -1 | -4 | -1 | -6 | -6 | -6 |

**Table 3: Relative difference between the widths of the quasi stochastic and stochastic 67% forecast intervals for population flows.**

| | Births. | Deaths | Net Migration |
|---|---|---|---|
| | % | | |
| yearly flow | | | |
| *d*(0-24 years) | -10 | -22 | -26 |
| *d*(25-47 years) | -7 | -50 | -57 |
| cumulated flow | | | |
| *d*(0-24 years) | 2 | -2 | -1 |
| *d*(25-47 years) | 2 | -6 | -1 |

# 5. Discussion

The results shows that it is possible to reproduce quite closely the main outcomes of a stochastic population forecast using only 6 variants. This can be done without fitting to the output of the stochastic forecast. Working with such a small number of variants does come at a price: one does not obtain a full probability distribution of the output variables, only a high and low forecast margin for a given confidence value. Also, forecast intervals for flows cannot be calculated. Intervals for cumulated flows can be obtained, however.

The difference between stochastic and quasi stochastic forecast intervals is larger for smaller age groups. This is mainly due to the way uncertainty in migration is modelled. The high and low Mig-variant give a consistent description of the uncertainty in cumulated net migration, but not for net migration in individual years. The gap between the high and low trajectories for net migration in the mig-variants increases in the first years of the forecast, but decreases in later years due to the narrowing factor. This means that, in the quasi stochastic method, more uncertainty is attributed to the size of the migration cohorts in the first few years of the forecast than in later years. This causes a distortion of the age profile of the forecast intervals that moves to older ages further on into the forecast. For large age groups, this distortion is averaged out to a greater extent than for small age groups.

As was discussed at the end of section 3.1, the quasi stochastic intervals for mortality-related uncertainty are expected to be too narrow, because the derivation of the high and low mortality variants did not take into account that input errors in the mortality rates have an effect on the population-at-risk of death (the elderly) that dampens the effects of these errors on the number of deaths in the subsequent years. For a long term forecast (25-48 years), it is found that the quasi stochastic interval for the number of elderly is too narrow by 6%, as is the interval for the cumulated number of deaths.

In section 3.2, it was assumed that there was no interaction effect between input errors in different components. This allowed the forecast intervals to be calculated from 6 variants instead of 8. To check this assumption, the quasi stochastic intervals were also calculated using these 8 combination-variants. The differences were found to be very small, with the highest effect, for green pressure, at 0.6% narrower intervals in 2060 for the 8-variant approach. The difference is so small not only because the interaction effects themselves are not large, but also because positive and negative interaction effects in the different combination-variants cancel each other in the calculation of the uncertainty interval using eq.(18).

# 6. Conclusions

Statistical institutes that compute variants for their population forecast, can, with little additional effort, compute forecast intervals for the main demographic indicators that are very similar to those obtained from a stochastic approach, at least for a forecast of up to 50 years into the future.

If the aim of the variants is to give a good description of the uncertainty in stock indicators like population size and grey pressure, the high and low variants for each component should aim to reproduce not the uncertainty in the yearly value of that component, but the uncertainty in the cumulated value of that component from the starting year of the forecast up to the forecast year under consideration. Equation (11) or (12) can be used to compute a narrowing factor to transform the first kind of variant into the second kind of variant.

Using equations (18) and (21), a quasi stochastic forecast interval can be calculated from the output of 6 of these variants, 2 for fertility, 2 for mortality, 2 for net migration.

Unlike a real stochastic forecast, this approach does not provide a full probability distribution for the forecast results. Also, it cannot be used to compute intervals for demographic flows, but intervals for time-cumulated flows can be obtained. The agreement with the stochastic intervals is better for larger age groups and for shorter forecast horizons. Also, it is better for population counts than for ratios of counts. The method works equally well for 67% and 95% forecast intervals.

An area where this approach can be useful is in more complex forecasts with many degrees of freedom or time consuming matching mechanisms, like subnational forecasts which employ housing market models. For this category of models, a stochastic forecast can be unfeasible, because it takes too long to run the simulations thousands of times. The quasi stochastic method can then be used to estimate approximate forecast intervals from a limited number of variants. The method was used in this way for the 2016-based regional population and household forecast of Statistics Netherlands and the Netherlands environmental assessment agency (Kooiman et al., 2016).

Acknowledgements: I would like to thank Nico Keilman and Joop de Beer for useful suggestions.

# 7. References


Alho, J. and B.D. Spencer (1985) Uncertain population forecasting, Journal of the American Statistical Association, 80, pp. 306-314.

Alho, J. and B.D. Spencer (1991) A population forecast as a database: implementing the stochastic propagation of error, Journal of Official Statistics, 7(3), pp. 295-310.

Alho, J.M. (1998) A stochastic forecast of the population of Finland. Reviews 1998/4, Statistics Finland, Helsinki.

Alders, M. and J. de Beer *1998) Kansverdeling van de bevolkingsprognose (Probability distribution of the population forecast). Maandstatistiek van de bevolking 46 (4), 8-11.

Alders, M., N. Keilman, J. Alho, T. Nikander (2005), Changing Population of Europe: Uncertain Future, UPE final report, European Commission, Report EUR 21699 EN.

De Beer, J. and M. Alders (1999) Probabilistic population and household forecasts for the Netherlands, Paper for the European Population Conference EPC99, The Hague, The Netherlands.

Duin, C. van and L. Stoeldraijer (2013) Bevolkingsprognose 2012-2060: langer leven, langer werken (Population forecast 2012-2060: living longer, working longer), Bevolkingstrends november 2013.

Hanika, A., W. Lutz and S. Scherbov (1997) Ein probabilistischer Ansatz zur Bevölkerungsvorausschätzung für Österreich. Statistische Nachrichten, 984-988.

Hyppölä, J., A. Tunkelo and L. Törnqvist (1949) Calculations on the population of Finland, its renewal, and its future development. Tilastollisia Tiedonantoja 38. Helsinki: Central Statistical Office of Finland (in Finnish).

Keilman, N. (2001) Uncertain population forecasts, Nature , ISSN 0028-0836. 412, pp. 490- 491

Keyfitz (1972) On future population, Journal of the American Statistical Association, 67 (338), pp. 347-363.

Kooiman, N, A. de Jong, C. Huisman, C. van Duin and L. Stoeldraijer (2016) PBL/CBS Regionale bevolkings- en huishoudensprognose 2016-2040: sterke regionale verschillen (Regional population and household forecast 2016-2040: strong regional differences), Bevolkingstrends 2016|08.

Lee, R. and S. Tuljapurkar (1994) Stochastic population forecasts for the United States: Beyond high, medium and low, Journal of the American Statistical Association, 89(428), pp. 1175–1189.

Lutz, W. and S. Scherbov (1998a) Probabilistische Bevölkerungsprognosen für Deutschland. Zeitschrift für Bevölkerungswissenschaft 23, 83-109/

Lutz, W. and S. Scherbov (1998b) An expert-based framework for probabilistic national population projections: the example of Austria. European Journal of Population 14, 1-17.

Pflaumer, P. (1998) Confidence intervals for population projections based on Monte Carlo methods, International Journal of Forecasting, 4, pp. 135-142.

Pötzsch ,O. and F.Rößger (2015) Bevölkerung Deutschlands bis 2060,13. koordinierte Bevölkerungsvorausberechnung, Statistisches Bundesamt, Wiesbaden, 2015.

United Nations (2014) World Population Prospects: The 2012 Revision, Methodology of the United Nations, Population Estimates and Projections, Working Paper No. ESA/P/WP.235.

# Appendix: Narrowing factor for an AR(1) model

The narrowing factor for the high migration variant is given by (eq.(11))

$$\psi_{\text{high}}(t) = \frac{\mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0}\Delta N(t_0+k)\right] - \mathrm{P}_{5/6}\left[\sum_{k=1}^{t-t_0-1}\Delta N(t_0+k)\right]}{\mathrm{P}_{5/6}\left[\Delta N(t)\right]}. \tag{A1}$$

For a normal distribution with mean zero, $\mathrm{P}_{5/6}$ and $\mathrm{P}_{1/6}$ are given by

$$\mathrm{P}_{5/6}[Y] = 0.967 \cdot \sqrt{\mathrm{E}[Y^2]}; \mathrm{P}_{1/6}[Y] = -0.967 \cdot \sqrt{\mathrm{E}[Y^2]}. \tag{A2}$$

where E[Y] denotes the expectation value of Y over its distribution.

Since the error terms in the time series (4) were assumed to have a symmetric normal distribution, the same holds for $\Delta N(t)$ and for sums of $\Delta N(t)$. Equation (A1) can therefore be written in the form

$$\psi_{\text{high}}(t) = \frac{\sqrt{E\left[\left(\sum_{k=1}^{t-t_0}\Delta N(t_0+k)\right)^2\right]} - \sqrt{E\left[\left(\sum_{k=1}^{t-t_0-1}\Delta N(t_0+k)\right)^2\right]}}{\sqrt{E\left[\Delta N(t)^2\right]}}. \tag{A3}$$

It follows from (A2) and (A1) that $\psi_{low}(t)$ and $\psi_{high}(t)$ are the same. Therefore, the label 'high' is dropped in the rest of the calculation.

Assuming that $\Delta N(t)$ follows an AR(1) time series model (4), the input error in net migration in forecast year *t* can be written as

$$\begin{aligned}\Delta N(t) &= \varphi^{t-t_0-1}\varepsilon(t_0+1) + \varphi^{t-t_0-2}\varepsilon(t_0+2) + \ldots + \varphi\varepsilon(t-1) + \varepsilon(t) \\ &= \sum_{k=0}^{t-t_0-1}\varphi^k\varepsilon(t-k).\end{aligned} \tag{A4}$$

From which it follows that

$$\begin{aligned}\mathrm{E}\left[\Delta N(t)^2\right] &= \mathrm{E}\left[\left(\sum_{k=0}^{t-t_0-1}\varphi^k\varepsilon(t-k)\right)^2\right] \\ &= \sum_{k=0}^{t-t_0-1}\sum_{k'=0}^{t-t_0-1}\varphi^k\varphi^{k'}\,\mathrm{E}\left[\varepsilon(t_0-k)\varepsilon(t_0-k')\right] \\ &= \sigma_N{}^2\sum_{k=0}^{t-t_0-1}\varphi^{2k}.\end{aligned} \tag{A5}$$

The cumulated input error can be written in the form

$$\begin{aligned}\sum_{k=1}^{t-t_0}\Delta N(t_0+k) &= \sum_{k=1}^{t-t_0}\sum_{l=0}^{k-1}\varphi^{l}\varepsilon(t_0+k-l)\\ &= \sum_{j=1}^{t-t_0}\varepsilon(t_0+j)\sum_{k=j}^{t-t_0}\varphi^{k-j},\end{aligned} \tag{A6}$$

where the terms in the summation have been reordered in the second step. It follows that

$$\begin{aligned}&\mathrm{E}\left[\left(\sum_{k=1}^{t-t_0}\Delta N(t_0+k)\right)^2\right]\\ &= \mathrm{E}\left[\left(\sum_{j=1}^{t-t_0}\varepsilon(t_0+j)\sum_{k=j}^{t-t_0}\varphi^{k-j}\right)^2\right]\\ &= \sum_{j=1}^{t-t_0}\sum_{j'=1}^{t-t_0}\sum_{k=j}^{t-t_0}\sum_{k'=j'}^{t-t_0}\varphi^{k-j}\varphi^{k'-j'}\,\mathrm{E}\left[\varepsilon(t_0+j)\varepsilon(t_0+j')\right]\\ &= {\sigma_N}^2\sum_{j=1}^{t-t_0}\left(\sum_{k=j}^{t-t_0}\varphi^{k-j}\right)^2.\end{aligned} \tag{A7}$$

Inserting this into (A3), the following expression is obtained for the narrowing factor for input errors generated by an AR(1) model with a symmetric normal distribution for the error term.

$$\psi(t) = \frac{\sqrt{\sum_{j=1}^{t-t_0}\left(\sum_{k=j}^{t-t_0}\varphi^{k-j}\right)^2} - \sqrt{\sum_{j=1}^{t-t_0-1}\left(\sum_{k=j}^{t-t_0-1}\varphi^{k-j}\right)^2}}{\sqrt{\sum_{k=0}^{t-t_0-1}\varphi^{2k}}}. \tag{A8}$$

Notice that the standard deviation has dropped out of (A8). This means that the factor does not depend on the magnitude of the input errors, only on their degree of autocorrelation.

## Explanation of symbols

| | |
|---|---|
| Empty cell | Figure not applicable |
| . | Figure is unknown, insufficiently reliable or confidential |
| * | Provisional figure |
| ** | Revised provisional figure |
| 2014–2015 | 2014 to 2015 inclusive |
| 2014/2015 | Average for 2014 to 2015 inclusive |
| 2014/'15 | Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015 |
| 2012/'13–2014/'15 | Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive |

Due to rounding, some totals may not correspond to the sum of the separate figures.

## Colofon

*Publisher*
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

*Prepress*
Statistics Netherlands, Studio BCO

*Design*
Edenspiekermann

*Information*
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information