WP 40
26 April 2010

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint Eurostat/UNECE Work Session on Demographic Projections
(28-30 April 2010, Lisbon, Portugal)

Item 13 – Stochastic national demographics projections

Developing stochastic population forecasts for the United Kingdom: Progress report and plans for future work

Steve Rowan and Emma Wright, Office for National Statistics, United Kingdom

ABSTRACT

The UK's Office for National Statistics (ONS) produces national population projections for the United Kingdom every two years. In addition to the principal (main) projections, variant projections are also produced. The variant projections are intended as plausible alternative scenarios and do not represent upper and lower limits for future demographic behaviour. In these variants, the different fertility, mortality and migration assumptions are treated as separate and independent departures from the assumptions in the principal projection.

One of the limitations of the traditional deterministic approach used to produce the national population projections is that no probabilities are attached to the projections; users are therefore given no information about the uncertainty associated with them. ONS is addressing this issue by developing a stochastic forecasting model for the United Kingdom.

This paper outlines the progress to date on developing a stochastic model. It describes how uncertainty about future demographic behaviour has been taken into account by expressing fertility, mortality and migration assumptions in terms of their assumed probability distributions. In each case, the median of the probability distribution is designed to follow the principal assumption from the 2006-based traditional deterministic projections. Three approaches for determining the probability distributions are also discussed. The paper reports the early findings of the research to date, and outlines ONS's plans for future work in this area.

It should be noted that the research reported is still in progress and, as such, any results presented are provisional only.

1. INTRODUCTION

National population projections (NPP) for the United Kingdom (UK) and its constituent countries are produced by ONS every two years. These demographic-based projections are essential for national planning functions in a range of fields and feed into sub-national, household, labour force and marital status projections. They are dependent on a set of assumptions about future levels of fertility, mortality and migration, which are reviewed and revised for each projection round. Details of the latest (2008-based) projections are described in *Population Trends* [1] and in the latest national population projections reference volume.[2] The complete results of the 2008-based principal (main) and variant projections are obtainable from the ONS website.[3] However, the work reported in this paper was conducted prior to the release of the 2008-based projections (in October 2009) and therefore focuses on the previous, 2006-based, national population projections released in October 2007. [4]

[1] Wright E (2010). 2008-based national population projections for the UK and constituent countries. *Population Trends 139*. Available from 25 March 2010 at: http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=6303
[2] ONS (2010) National Population Projections 2008-based (PP2 no. 27).
Available from 25 March 2010 at: http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=4611
[3] Full results of the 2008-based national population projections available at: http://www.statistics.gov.uk/statbase/Product.asp?vlnk=8519
[4] Full results of the 2006-based national population projections, and previous UK population projections, available on Government Actuary's Department (GAD) website at: www.gad.gov.uk/Demography%20Data/Population/index.aspx

The principal population projections are based on assumptions considered to be the best that could be made at the time they are adopted. Variant projections are also produced, which are intended as plausible alternative scenarios and do not represent upper or lower limits for future demographic behaviour. In these variant projections, the different fertility, mortality and migration assumptions are treated as separate and independent departures from the assumptions in the principal projections. 'Single component' variants, varying just one of the three components, are produced, as well as selected 'combination' variants produced by combining two or more of these alternative scenarios.

It is increasingly being recognised, however, that the traditional deterministic approach has a number of limitations. As stated earlier, no probabilities are attached to the principal projections so users are given no information about the uncertainty associated with them or, with respect to the variants, are given no indication of how these compare to the principal projections in terms of certainty. In response to these concerns, increasing attention is now being given to stochastic forecasting methods. Typically, stochastic forecasts use probability distributions for the components of demographic change, namely of fertility, mortality and migration. These are derived using some combination of three recognised approaches: analysis of past projection errors, expert opinion and time series analysis. By using these approaches, ONS is developing a stochastic forecasting model for the United Kingdom.

This paper describes how uncertainty about future demographic behaviour has been taken into account by expressing fertility, mortality and migration in terms of their assumed probability distributions. In each case, the median of the probability distribution was designed to follow the principal assumption from the traditional deterministic projections in order to assess the uncertainty associated with the published deterministic projections.

It is important to recognise that stochastic forecasts have their limitations too. In particular, there is a risk that specifying precise probability ranges may convey a misleading sense of precision to users.[5] The true probability distribution for the future total fertility rate (TFR), for example, is not a known quantity. In fact, just as with deterministic projections, the validity of stochastic forecasts is dependent upon the accuracy of the assumptions underlying the model.

Population projections are subject to considerable uncertainty due to potential changes in a wide range of factors including the economy, the impact of government policies, individual, family and household behaviour and events external to the UK. Such explanatory variables (whether economic or not) may be important drivers of population change and may affect levels of fertility, mortality and migration. However, in the long-term, they are considered to be as difficult to predict, if not more so, than demographic variables. Bearing all this in mind, any set of projections will inevitably be proved inaccurate to a greater or lesser extent.

The following sections outline the work conducted to date. There remain a number of outstanding issues that need to be considered in future analyses and these are covered in the 'Future work' section.

2. COMPONENTS OF DEMOGRAPHIC CHANGE

The first stage of the research involved selecting the components of demographic change or 'drivers' for the model. For fertility, the total fertility rate (TFR) was selected (see Appendix A). Male and female period expectations of life at birth (EOLB) were chosen as the mortality drivers. For migration, total net migration was used. All the chosen drivers are standard indicators for demographic stochastic forecasting models. The TFR and net migration are also both directly used in setting the assumptions for the deterministic projections. Although EOLB is also used as a headline assumption indicator in the deterministic projections, the mortality assumptions are actually formulated in terms of expected rates of annual mortality improvements.

For international migration, it was decided to model net migration as opposed to immigration and emigration separately, which is consistent with the UK deterministic projection model. There are conceptual issues with this decision as there is no such thing as a net migrant. However, when analysing past projection errors, only total net migration data are available and so analysis of gross flows could not be carried out. Also, for the 2006-based deterministic projections, the NPP Expert Panel had only been asked for their views on future levels of net migration. The choice of net migration - as opposed to immigration and emigration separately, is discussed further in the 'Future work' section. Using numbers (as opposed to rates) for net migration is also consistent with the deterministic projections model and seemed sensible since only emigration has a clear base population on which to base rates.

[5] Lutz W and Goldstein JR (2004) Introduction: How to Deal with Uncertainty in Population Forecasting? *International Statistical Review*, Volume 72, Number 1. Available at: http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.isr/1079360109

3. DERIVING PROBABILITY DISTRIBUTIONS FOR THE COMPONENTS OF DEMOGRAPHIC CHANGE

There are three widely recognised approaches for determining the probability distribution for each driver: past projection error, expert opinion and time series analysis. When determining the driver distributions, all three approaches were considered.

3.1 Past projection error

One of the key ways of indicating uncertainty is to consider the accuracy of past sets of projections by comparing historical projections against actual numbers taken from the published series of population estimates. Analysis of past projection error not only tells users of the accuracy of past projections but, if past accuracy is a reliable guide to future accuracy, can also be used to inform on the likely magnitude of errors in future projections. A detailed study of the accuracy of past UK national population projections was published in 2007[6] and was based on the extensive database of past national projections which is available on the GAD website.

Similarly, the accuracy of the fertility, mortality and migration assumptions from past sets of projections can be compared with actual values. Using the GAD database, 18 sets of projections from the 1971-based projections up to the 2004-based projections were used to identify past projection errors. For each driver, the root mean squared error (RMSE, see Appendix A) was calculated for one year ahead and for all years up to twenty five years ahead. The RMSE provides a single measure of the difference between the original assumptions and actual values at any time into the projection period.

Over the years there have been many population revisions and it is predominantly migration that has been affected. However, no adjustment has been made to allow for such revisions in the calculation of RMSEs. If the first few projection years are ignored, as these will be the most affected, the RMSEs for five years ahead durations (and upwards) are much more reliable indicators of 'real' projection error.

3.2 Expert opinion

Expert views on future demographic behaviour have been obtained from the NPP expert advisory panel.[7] They provide advice on the appropriate assumptions to use for the NPP. In 2007, ahead of the 2006-based projections, the six experts on the panel were asked to complete a detailed questionnaire in which they were asked for their opinions on the most likely levels (with 67 per cent confidence intervals) for the TFR, male and female EOLB and annual net migration in 2010 and 2030, which was five and twenty five years ahead of the latest data available at the time.

For each of the drivers, a mean of the six responses and a mean of both the upper and lower confidence intervals (CIs) were calculated. Whenever an interval was non-symmetric, the wider interval was chosen to ensure symmetry. An assumption was made that for each driver, these 'average' confidence intervals came from a normal distribution with mean equal to the 'average' most likely level. From this, standard deviations were estimated and 95 per cent confidence intervals for five and twenty five years ahead were derived.

3.3 Time series

The time series method of analysis uses the trends and variability of historical time series of fertility, mortality and migration indicators to produce estimates of the standard deviations around future values. The time series models explored included simple exponential smoothing, log linear (Holt) exponential smoothing and linear trends. In general, the resultant probability distributions differed considerably from those distributions derived using past projection error and expert opinion.

Consequently, time series techniques have not been used in this early research to determine the probability distributions described in this paper. However, it is recognised the use of a time series approach needs to be explored more fully in the future (see 'Future work' section).

[6] Shaw C (2007). Fifty years of United Kingdom national population projections: how accurate have they been? *Population Trends 128*. pp 8-23. Available at: http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=6303

[7] Details of the most recent meeting of the NPP Expert Advisory Panel in spring 2009 are available at http://www.statistics.gov.uk/downloads/theme_population/NPP2008/NatPopProj2008.pdf (see section 11). Details of the meeting of the panel held prior to the production of the 2006-based projections are available at: http://www.gad.gov.uk/Demography Data/Population/2006/methodology/expert1.html

4. COMPARISON OF PAST ACCURACY AND EXPERT OPINION APPROACHES

The five and twenty five years ahead standard deviations derived from expert opinion are compared with RMSEs derived from past projection errors in Table A.

Table A - Comparative measures of uncertainty for five and twenty five years ahead

		TFR (number of children)	male EOLB (years)	female EOLB (years)	Net migration (000s)
Five years ahead	Experts: standard deviation	0.15	1.18	0.79	**51.3**
	Past accuracy: RMSE	0.20	0.78	0.66	58.6
Twenty five years ahead	Experts: standard deviation	**0.26**	**2.31**	**2.05**	90.7
	Past accuracy: RMSE	0.50	4.79	3.30	162.4

Note: The bold values were used to derive the sample paths for the drivers used in the final model (see 'Forecast assumptions' section).

Table A shows the standard deviations derived from expert opinion and the RMSEs derived from past projection error at both five and twenty five years ahead. Although some exploratory work has been carried out, more work is required to test whether differences can be attributed to sampling error by calculating uncertainty around each standard deviation or RMSE. This is particularly important for expert opinion given that the results are based on a small sample of six experts. The accuracy of the estimated standard deviations will also be dependant on the underlying assumption that the experts' 'average' confidence intervals are taken from a Normal distribution. This is discussed further in the 'Future work' section.

For past projection error, no distribution has been assumed. The direction of the error over time and any bias towards under or overestimation, are discussed in the 'Forecast assumptions' section.

5. THE MODEL

The overall model uses a version of the cohort component model used to produce the deterministic projections, as follows:

$$P_t = P_{t-1} + B_t - D_t + M_t$$

where *t* denotes the year of forecast, P_{t-1} is the level of the population in the previous year and B_t, D_t and M_t are total counts of births, deaths and net migration, respectively.

The difference is that instead of specifying fertility, mortality and migration for the whole projection period, this model uses counts of births, deaths and net migration derived from the stochastically determined drivers of fertility, mortality and migration. (The 'Sample path' section describes how total counts of births, deaths and net migration were derived.)

The probability distributions for year to year change are generated by carrying out a large number of simulations (or sample paths) of the overall model.

A random walk with drift (RWD) model was chosen to generate the drivers for each sample path. The level of each driver from one year to the next is:

$$Driver_t = Driver_{t-1} + Value_t + Drift_t$$

where *t* denotes the year of forecast, $Driver_{t-1}$ is the level of the driver in the previous year, $Value_t$ is a random value sampled from a normal distribution of year on year differences with a mean of zero and a pre-selected one-year ahead

standard deviation. Details of the one year ahead standard deviations chosen for each driver are given in the 'Forecast assumptions' section. $Drift_t$ is a drift term which constrains the median of the probability distribution to follow the 2006-based UK principal assumptions. The drift term is calculated as:

$$Drift_t = PDriver_t - PDriver_{t-1}$$

where $PDriver_t$ and $PDriver_{t-1}$ are driver values taken directly from the 2006-based UK principal assumptions.

To get the random value, a random number is generated from a normal distribution using a Box Muller[8] approach. Each random number is then multiplied by a pre-selected one year ahead standard deviation (which, for each driver, is fixed throughout the projection period).

The one year ahead standard deviation is calculated as follows:

$$SD_1 = SD_t / \sqrt{t}$$

where SD_1 is the unknown standard deviation at one year ahead; and

SD_t is the known standard deviation at t years ahead (derived from past projection error or expert opinion).

Making the assumption of perfect correlation between the sexes, the same random numbers were used for male and female EOLB. However, for each of the three different drivers, the random numbers were generated separately. This implicitly assumes the probability distributions for each driver are independent of each other. This does not imply that the number of births and deaths are unaffected by changes in population size due to migration.

6. FORECAST ASSUMPTIONS

6.1 Fertility

The TFR sample paths were generated using a RWD model where the drift term constrained the median of the probability distribution to follow the TFR principal assumption from the 2006-based UK NPP, namely a long-term TFR of 1.84.

In order for the TFR sample paths to have a probability distribution that closely matched the experts' 'average' 67 per cent CI at twenty five years ahead, the derived experts' standard deviation of ±0.26 children was chosen (see Table A). A one year ahead standard deviation of 0.05 children was calculated and fed into the random value component (where $0.05 = 0.26/\sqrt{25}$).

Figure 1 shows that the 67 per cent prediction intervals broadly match the high/ low fertility variants of the NPP, but are much narrower than the experts' 'average' 67 per cent CIs at five years ahead. However, the 95 per cent prediction intervals closely match the derived experts' 95 per cent CI at twenty five years ahead. The RMSEs (based on past projection error) are also included in the chart for information.

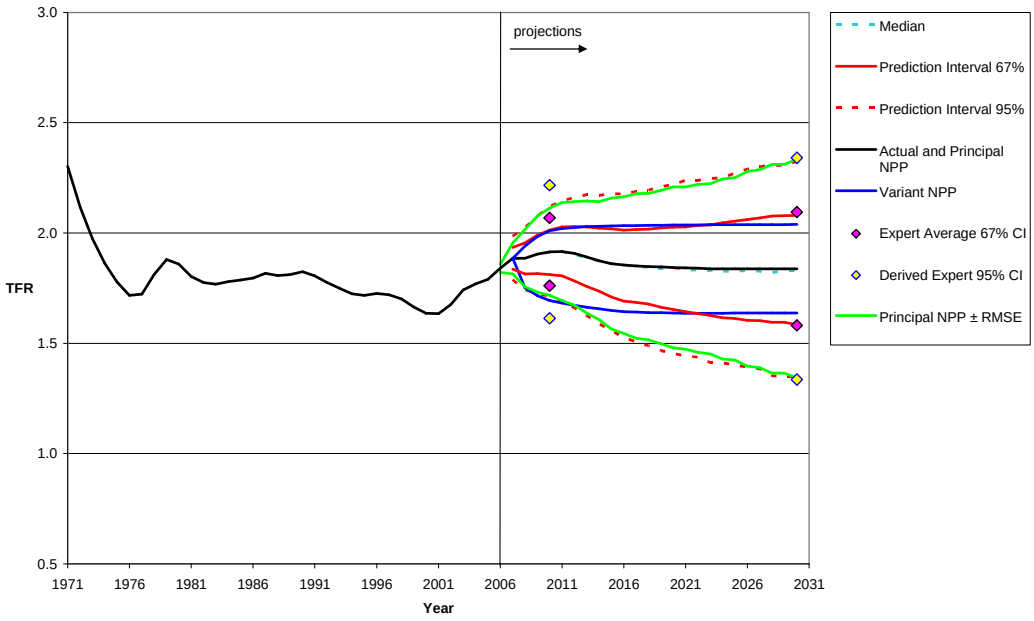
In describing the traditional variants as "plausible alternatives" rather than limits, experts[9] have assumed the TFR range covered by the high and low variant assumptions is approximately 70 per cent and Figure 1 gives some support for this in the short-term (up to 2030). However, since the long-term TFR is held constant in the principal and variant deterministic projections, the gap between the variants and the prediction intervals will widen over time.

Table A shows that the RMSEs for five and twenty five years ahead are higher than the standard deviations estimated from expert opinion. The past projection error encompasses some projections (beginning with 1971-based) made following the 1960s baby boom when fertility levels were falling rapidly and errors were unusually high. In general, for all the projection sets, the past projection errors were positive and increased in magnitude as time from the base year progressed, indicating a systematic past bias towards overestimation of fertility.

[8] Golder E and Settle J (1976) The Box-Muller Method for Generating Pseudo-Random Normal Deviates. *Applied Statistics*, Vol. 25, No. 1, pp. 12-20

[9] Scherbov S, Mamolo M, Lutz W, Probabilistic Population Projections for the 27 EU Member States Based on Eurostat Assumptions. Available at: http://www.oeaw.ac.at/vid/download/edrp_2_08.pdf

Figure 1. 2006-based TFR: Median and prediction intervals compared to the principal projections, variants, experts' CIs and principal NPP ± RMSE, 1971-2030, United Kingdom



6.2 Mortality

For each sex, the EOLB sample paths were generated using a RWD model where the drift term constrained the median of the probability distribution to follow the EOLB principal assumption from the 2006-based UK NPP, namely that the EOLB in 2031 is 82.7 years for males and 86.2 years for females.

In order for the EOLB sample paths to have a probability distribution that closely matched the experts' 'average' 67 per cent CIs at twenty five years ahead, the derived experts' standard deviation of ±2.31 years for males and ±2.05 years for females was chosen (see Table A). A one year ahead standard deviation of 0.46 years for males and 0.41 years for females was calculated and fed into the random value component (where $0.46=2.31/\sqrt{25}$ and $0.41=2.05/\sqrt{25}$).

Figures 2 and 3 show that the 67 per cent prediction intervals are wider than the high/ low life expectancy variants of the NPP, but become closer towards the end of the projection period shown. The RMSEs (based on past projection error) are also included in the charts for information.

For EOLB, the past projection errors were negative and increased in magnitude as time from the base year progressed, indicating a systematic past bias towards underestimation of life expectancy.

It should be noted that year on year change is assumed to be perfectly correlated for males and females. In other words, for any given sample path, the random value terms (having adjusted for the slightly greater standard deviation for males than females) are the same. But in practice, year on year change in male and female life expectancy is strongly rather than perfectly correlated.

Figure 2. 2006-based EOLB Males: Median and prediction intervals compared to the principal projections, variants, experts' CIs and principal NPP ± RMSE, 1971-2030, United Kingdom

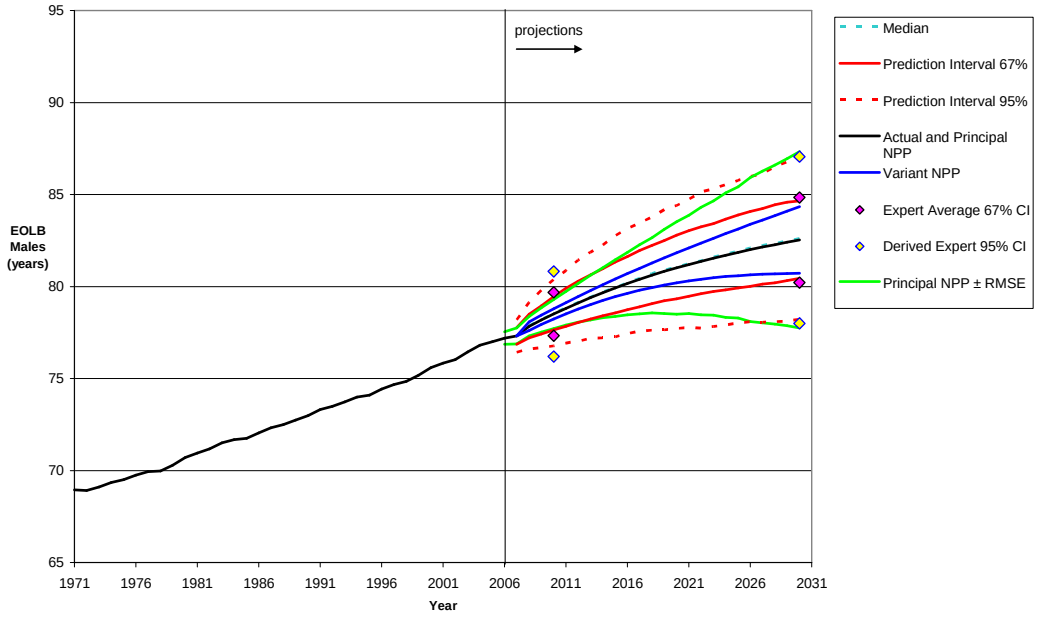
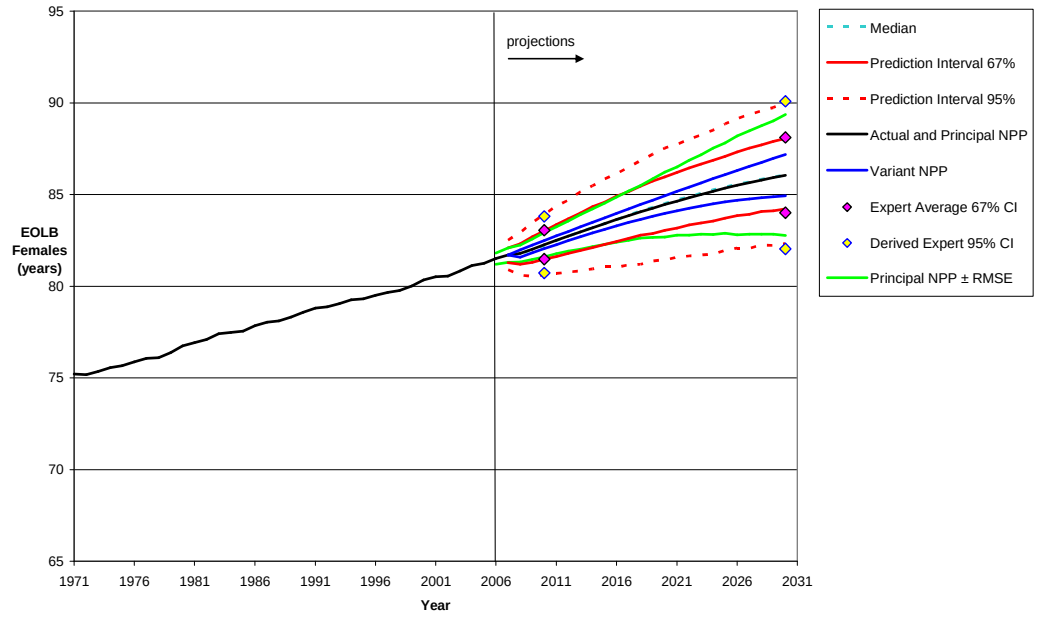


Figure 3. 2006-based EOLB Females: Median and prediction intervals compared to the principal projections, variants, experts' CIs and principal NPP ± RMSE, 1971-2030, United Kingdom



6.3 Migration

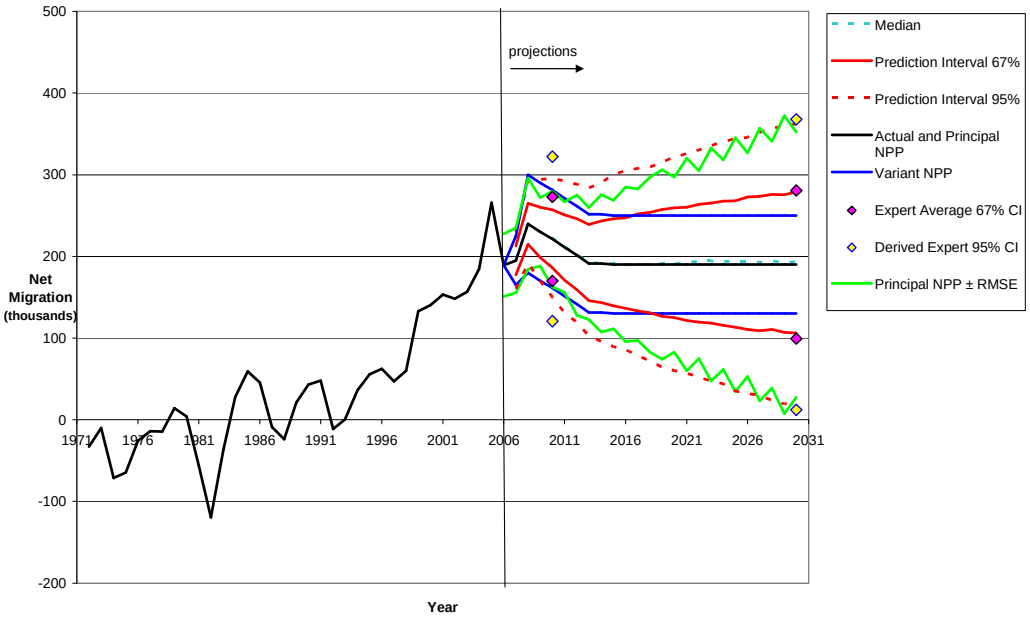
The net migration sample paths were generated by a RWD model where the drift term constrained the median of the probability distribution to follow the net migration principal assumption from the 2006-based UK NPP, namely a long-term value of +190,000 net migrants per year. Two alternatives for the one year ahead standard deviation were considered.

A) In order for the net migration sample paths to have a probability distribution broadly consistent with the experts' 'average' 67 per cent CI at twenty five years ahead, the derived experts' standard deviation of ±90,700 net migrants was

chosen (see Table A). A one year ahead standard deviation of 18,100 net migrants was calculated and fed into the random value component (where 18,100=90,700/$\sqrt{25}$).

Figure 4 shows that the 67 per cent prediction intervals are narrower than the high/ low net migration variants of the NPP in the short-term but wider in the longer term; this can in part be explained by the fact that in the deterministic projections, long-term net migration levels are held constant and therefore do not assume a continuous level of uncertainty. The 67 per cent prediction intervals are much narrower than the experts' 'average' 67 per cent CI at five years ahead. However, the 95 per cent prediction intervals closely match the derived experts' 95 per cent CI at twenty five years ahead. The RMSEs (based on past projection error) are also included in the chart for information. Note: the jagged appearance of the principal NPP ± RMSEs lines are due to a larger error for the year 2004 (in all projection sets) relative to earlier years.

Figure 4. 2006-based Net Migration: Median and prediction intervals compared to the principal projections, variants, experts' CIs and principal NPP ± RMSE, 1971-2030 (with one year ahead standard deviation set at 18,100), United Kingdom

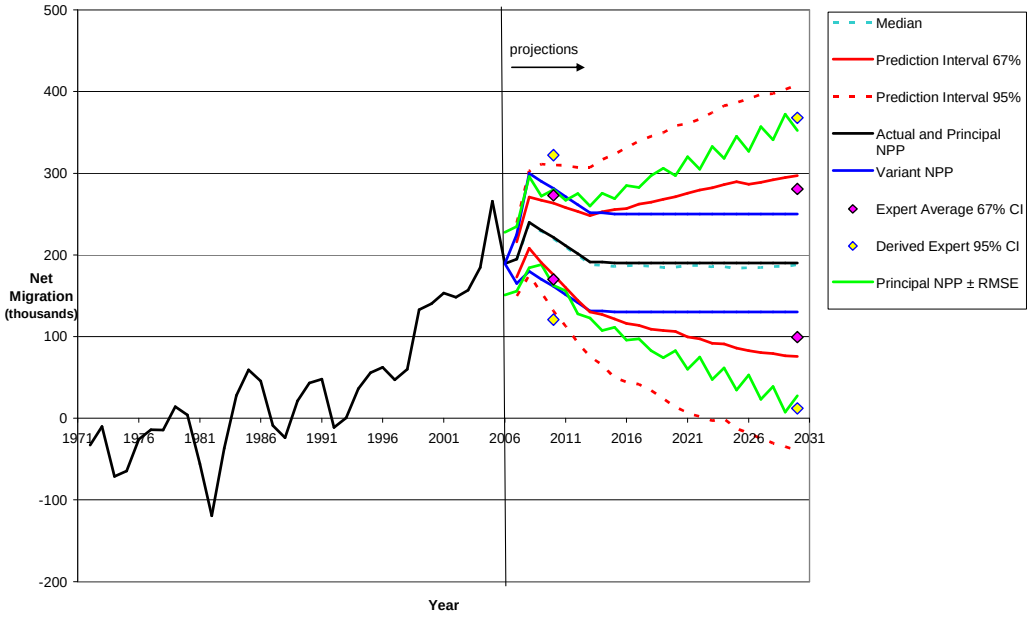


For net migration, the past projection errors were negative and increased in magnitude as time from the base year progressed, indicating a systematic past bias towards underestimation of net migration.

B) In order to ensure a probability distribution broadly consistent with the experts' 'average' 67 per cent CI at five years ahead, the derived experts' standard deviation of ± 51,300 net migrants was chosen (see Table A). A one year ahead standard deviation of 23,000 net migrants was calculated and fed into the random value component (where 23,000=51,300/$\sqrt{5}$).

Figure 5 shows that the 67 per cent prediction intervals are wider than the experts' 'average' 67 per cent CI at twenty five years ahead.

Figure 5. 2006-based Net Migration: Median and prediction intervals compared to the principal projections, variants, experts' CIs and principal NPP ± RMSE, 1971-2030 (with one year ahead standard deviation set at 23,000) United Kingdom



As Figures 4 and 5 illustrate, neither set of probability distributions provides a particularly good fit to the experts' 'average' 67 per cent CIs in both the short-term and longer term. It was decided therefore, to use a one year ahead standard deviation of 23,000 (as shown in Figure 5) to generate sample paths for net migration because it was deemed preferable to overestimate rather than underestimate uncertainty.

7. SAMPLE PATH

This section describes the process of how the stochastically determined drivers were converted into counts that fed into the overall model. This process represents one sample path or simulation. A large number of simulations were needed in order to obtain the required probability distributions.

7.1 Fertility

For every projection year, the TFR was derived using a RWD model as described in 'The model' section. An age distribution by single year of age of mother (ranging from 15 to 46 years old) was derived using the 2007 age-specific fertility rates (ASFR) for the UK. This (fixed) age distribution was applied to the stochastically determined TFRs to give ASFRs for subsequent projection years. See Appendix A for the formulae for ASFRs and age distribution.

Applying the ASFRs to the female population gave a count of births by single year of age of mother. The births were summed to give a count of total births which were then fed into the overall model. To obtain total births by sex required applying the male: female sex ratio of 51.2 : 48.8 to total births.

7.2 Mortality

For every projection year (and separately by sex), EOLBs were derived using a RWD model as described in 'The model' section. Each stochastically determined EOLB was fed into a look up procedure that picked out the associated 2006 age-specific mortality rate from a model life table based on historic life tables and the 2006-based UK principal mortality assumptions.

A count of deaths by age was generated by applying the relevant age-specific mortality rates to the population (adjusted for net migration). For children not yet born, the given mortality rate (from the model life table) was applied to total births to give counts of infant deaths. The age-specific counts of deaths were summed to give total deaths that were fed into the overall model.

7.3 Migration

For every projection year, total net migration was derived using a RWD model as described in 'The model' section. The counts of total net migration were fed into the overall model.

However, as stochastic forecasts by sex and age are required, a number of additional steps were needed to derive age/ sex-specific net migration:

- It was necessary to split net migration into immigration and emigration. Total emigration was fixed to the level of the 2006-based UK principal migration assumptions (475,000 in the long-term) and then total immigration was calculated from the stochastically determined net migration. It is planned to revisit this assumption of a fixed level of emigration.
- Age and sex distributions (different for emigration and immigration), were also taken from the 2006-based UK principal migration assumptions. For each age and sex, net migration was then calculated as the difference between immigration and emigration.

In the deterministic projections, it is assumed that levels of annual net migration beyond 2015 will remain constant. In reality, there will be fluctuations from year to year, but these are very difficult to predict. Short-term assumptions have been applied to the first few years of the deterministic projections and this includes allowance for additional net migration from accession countries.

Starting population

The mid-2006 UK population estimates published in August 2007[10] were used as the starting population. These are the same estimates used in the 2006-based UK NPP.

8. THE PROGRAM

The program was written in Microsoft Excel and used an adapted version of the cohort component model used to produce the deterministic projections. For each of the drivers, the value one year ahead was obtained by adding the stochastically determined year on year change and a drift term to the level of the driver in the previous year. The inclusion of the drift term ensured the median of the assumed probability distributions was consistent with the results of the 2006-based UK principal assumptions. However, some very small differences arose due to the way age-specific rates and numbers were obtained.

For each driver, (but simultaneously with the other drivers) a new simulation or sample path was created each time a new set of random numbers was generated. The resultant values were then converted into counts (as described in the 'Sample path' section) and added to the total population counts from the previous year. The program was designed to generate 5,000 simulations from 2006 through to the year 2056.

Illustrative results

Figures 6 and 7 below show the provisional projected populations and the level of uncertainty at two points in the future (2031 and 2056) by age and sex using the assumptions described in this paper. They are for illustrative purposes only.

Figure 6 shows that, in absolute terms, uncertainty is greatest for the youngest cohorts - particularly those who are yet to be born. Migration will have an impact for those cohorts which pass through the peak ages of migration in the next 25 years. There is also uncertainty at the very oldest ages; however, while this uncertainty may be small in absolute terms, it is significant relative to the median population size for this age group.

The chart also shows that there is least uncertainty for those who are aged about 55-75 years old in 2031 - in other words, those who are currently aged around 30-50 years old. These cohorts are past the peak ages of migration and will not be significantly affected by mortality in the next 25 years.

[10] Mid-2006 Population Estimates for the UK, August 2007. Available at:
http://www.statistics.gov.uk/downloads/theme_population/Mid_2006_UK_England_&_Wales_Scotland_and_Northern_Ireland%20_22_08_07.zip

Figure 6. Provisional UK projected populations: Median and prediction intervals, 2031, United Kingdom

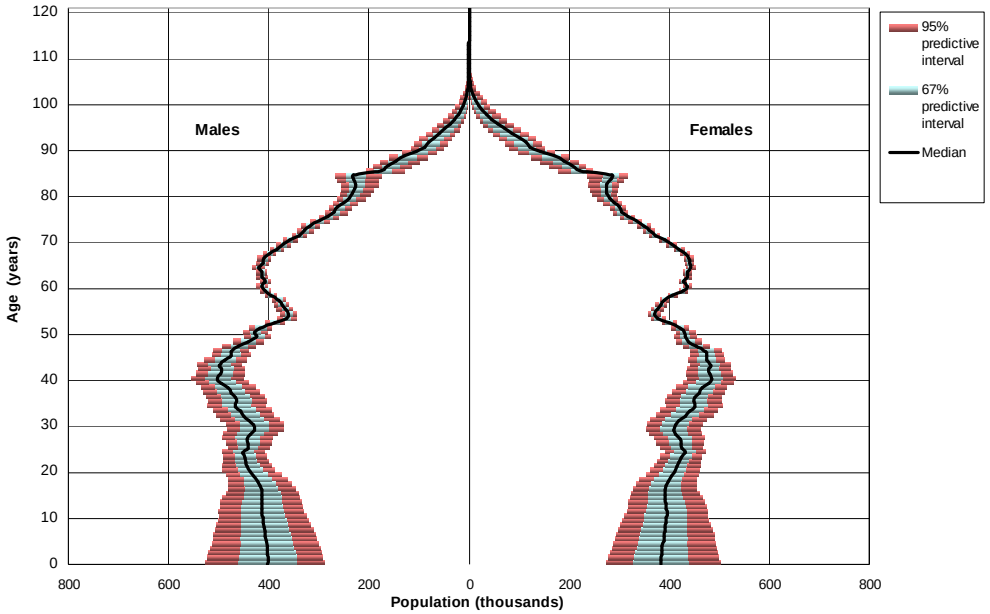
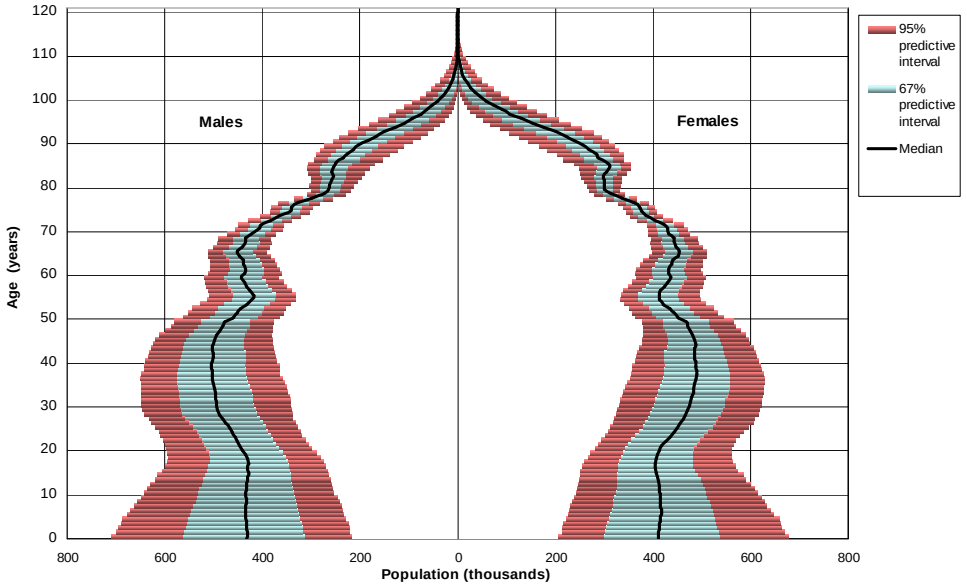


Figure 7 shows how uncertainty clearly increases 50 years into the future - particularly for cohorts yet to be born. Further into the projection period, uncertainty about the number of births is not only dependent on the TFR, but also because of the uncertainty about the size of future child bearing cohorts which will be affected by both fertility and migration uncertainty. The chart shows least uncertainty for those in their late 70s in 2056 (or those who are presently in their late 20s).

Figure 7. Provisional UK projected populations: Median and prediction intervals, 2056, United Kingdom



9. FUTURE WORK

9.1 Use of a time series approach

The research reported in this paper has been based on past projection error and expert opinion. However, time series models appear to have provided satisfactory results in other stochastic forecasting exercises.[11] ONS intends to reconsider the time series approach exploring, in particular, whether the most appropriate models were used. Consideration will also be given to fitting time series models using historical data series of different time spans and comparing the results. This is because results from time series analyses are very sensitive to the historical reference period considered. So, for example, if the period included the 1960s baby boom or the fertility decline in the early 1970s, then this could have an impact. If we believe the TFR will continue around its current level indefinitely, then fitting a time series model from the late 1970s would provide a good representation of observed variability around this level.

9.2 Deriving probability distributions

In this early work, it has been assumed that the 'average' 67 per cent confidence intervals for the drivers provided by the expert panel are from Normal distributions with mean equal to the 'average' most likely level. Yet many of the confidence intervals provided by individual experts were skewed. Hence this assumption needs further consideration.

The NPP panel were asked to provide both 67 per cent and 95 per cent confidence intervals for the 2008-based NPP and as a result, future distributional assumptions could have a more solid foundation. Future work should also consider the distribution of past projection errors.

9.3 The RWD model

While the RWD model may be appropriate to produce probability distributions for the drivers, rigorous time series analysis needs to be carried out for each driver to test this assumption, and to determine what the best model is based on the analysis of historical data. Fitting the RWD model to past data will not only provide a good test of the model, but will also provide an estimate of the variance of the random value term. If a RWD model is not appropriate, an ARIMA model could incorporate serial correlations for both assessing the underlying variance and making projections.

State Space or Kalman Filter models could also be considered. These use parameters from past data to predict the future. For every age and parameter there is a random adjustment and these are all correlated.

9.4 Correlations

Correlations between ages and between sexes (in both mortality and migration) need to be considered. The current methodology assumes that all ages are perfectly correlated (within a component) which may not be true. An analysis of the correlation between mortality, fertility and migration over time should also be considered and could be carried out using historical data. Expert opinion could also be sought or existing academic analysis considered.

9.5 Net Migration

Further work into the viability of modelling immigration and emigration separately should be considered as it is not only conceptually easier, but would increase the transparency of the assumptions. The NPP panel have provided views on emigration and immigration (in addition to net migration) for the 2008-based NPP. Estimates of the standard deviation for immigration and emigration separately over time could be obtained by examining historic data, and the feasibility of estimating synthetic RMSEs by comparing actual immigration, emigration and net migration from the population estimates with projected net migration could be explored.

In future analysis, consideration could be given to splitting migration flows further, for example, by country of origin and forecasting each flow separately. There are a number of different types of migrants including labour, family, student and asylum migrants suggesting the use of a fuller model with more drivers and possibly the inclusion of external (economic) factors may be more appropriate.

[11] Alders M., Keilman N. and Cruijsen H. (2007) Assumptions for long-term stochastic population forecasts in 18 European countries. *European Journal of Population*. Available at: http://www.springerlink.com/content/0453r268h24456q8/fulltext.pdf

9.6 Age distribution

The age distribution is fixed for fertility and mortality, and also for net migration from 2015. This means the overall model takes no account of possible change in the age distributions over time such as any continued increase in the average age of mother at birth for example, or any change in the long-term disparity between the ages of people leaving the UK compared to those entering. Therefore, the option of disaggregating by age should be explored.

9.7 The program

Future work will consider whether 5,000 simulations are sufficient, whether there is evidence of convergence and test for stability using a trace plot.

ACKNOWLEDGEMENTS

This paper reports work conducted by Cath Brand, Mita Saha, Chris Shaw and Steve Rowan from the ONS Centre for Demography. Many thanks to Professor Phil Rees from the University of Leeds, Professor Nico Keilman from the University of Oslo, Professor Wolfgang Lutz from the Vienna Institute of Demography and Ruth Fulton and colleagues from ONS Methodology Directorate for their advice and input.

APPENDIX A

Formulae

- The Root Mean Squared Error (RMSE) for *m* years ahead is given by

$$RMSE = \sqrt{((\sum_{t=1}^{N_m} E_{m,t}^2) / N_m)}$$

where $E_{m,t}$ = *m* years ahead forecast value from projection set *t* minus actual value

and N_m = number of sets of projections for *m* years ahead.

- The age-specific fertility rate (ASFR) in 2007 is given by

$$ASFR_{n,2007} = (B_{n,2007} / P^f_{n,2007}) * 1000$$

where B_n = live births to women at age *n*,

P^f_n = female population at age *n*,

and n = single year of age of mother, (15, …, 46).

- The Total Fertility Rate (TFR) in 2007 is given by

$$TFR_{2007} = (\sum_{n=15}^{n=46} ASFR_{n,2007}) / 1000$$

- The age distribution (AD) for fertility in 2007 is given by

$$AD_{n,2007} = ASFR_{n,2007} / \sum_{n=15}^{n=46} ASFR_{n,2007}$$

$$= ASFR_{n,2007} / TFR_{2007} * 1000$$

- ASFRs derived for a sample path are given by

$$ASFR_{n,t} = TFR_{n,t} * AD_{n,2007} * 1000$$

where *t* denotes the year of forecast.