Evaluating probabilistic demographic forecasts

Probabilistic forecasts give results either in terms of predictive probability distributions, or as simulated samples. Sometime after the forecast has been computed, the actual value of the variables in question can be compared with the distributions. With these observations, how can we evaluate the probabilistic forecast?

Statisticians have developed so-called scoring functions. A scoring function is a measure for the distance, defined in a specific way, between the distribution and the outcome. However, scoring functions are not widely known among demographers, in spite of the fact that more and more often demographic forecasts are computed as probability forecasts (e.g. the World Population Prospects by the UN Population Division).

The aim of the paper is to review scoring functions and their use in demographic applications. It turns out that they have been used in the calibration of some probabilistic demographic forecasts, based on hold-out samples as ex-post data. There are no known applications of scoring functions to probabilistic demographic forecasts computed several years ago. Thus, a second aim is to illustrate scoring functions by evaluating several probabilistic forecasts that became available since the end of the 1990s. For these forecasts, we have two decades of data.

We will evaluate probabilistic population forecasts published by Statistics Netherlands in 1998, and the UPE forecasts for selected European countries with jump-off year 2003. We will also evaluate a probabilistic household forecast. To compare scores for forecasts with different jump-off years, we need analyses based on age-period-cohort type of models.

1. Introduction

Deterministic forecasts do not quantify prediction uncertainty. They may contain a high variant and a low variant for population growth, in addition to the medium variant or most likely trajectory of future population growth. The variant for strong population growth combines a high assumption for future fertility with one on high life expectancy and high immigration. Weak population growth is based on low fertility, low life expectancy, and low immigration in the future. However, we do not know whether chances are 30%, or 60%, or 90% that actual population will be inside the interval determined by the low and the high variants.

This is why the statistical agencies of some countries have started to publish their forecasts in the form of probability distributions, following common practice in, for example, meteorology and economics. Statistics Netherlands pioneered the field; see Alders and De Beer (1998). Statistics New Zealand (2011) and Statistics Italy (ISTAT, 2018) are the other two known examples. In this connection, one should also mention the Population Division of the United Nations, which is responsible for regular updates of population forecasts for all countries of the world. In 2014, the Population Division issued the first official probabilistic population forecasts for all countries, using the methodology developed by Raftery et al. (2012); see also http://esa.un.org/unpd/wpp/Graphs/Probabilistic/POP/TOT/. The aim of a probabilistic forecast is *not* to present estimates of future trends that are more accurate than a deterministic forecast, but rather to give the user a more complete picture of prediction uncertainty.

Once a probabilistic forecast has been published, some 10-20 years later its accuracy can be evaluated, when *ex-post facto* observed data for population size and age structure have become available. However, accuracy assessment is difficult to carry out directly because it requires comparing a forecaster's predicted probabilities with the actual but unknown probabilities of the events under study. For that reason, statisticians have developed "scoring rules", which are empirical distance measures between the predicted distribution of the demographic variable in question, and the empirical value it actually turns out to have. Gneiting and Raftery (2007) and Gneiting and Katzfuss (2014) review the field. The score that one finds for a certain variable has no intrinsic meaning. Only in a comparative perspective, one can interpret the scores in a useful manner. Indeed, scoring functions are used in comparing two or more competing probabilistic forecasts.

Although the methodology around evaluation of probabilistic forecasts and scoring rules has been known for some time, there are very few applications of scoring rules to population forecasting. Shang et al. (2016) evaluated the accuracy of probabilistic cohort-component forecasts for the UK, and compared two forecasting methods. They used a scoring rule for prediction intervals. Shang and Hyndman (2017) evaluated interval forecasts for age-specific mortality rates of Japan, and used interval scores to select the best among three methods. Alexopoulos et al. (2018) employed interval scores to prediction intervals of age-specific mortality of England & Wales and New Zealand, and evaluated the predictive performance of five different mortality prediction models. All three papers use hold-out samples to evaluate the probabilistic demographic forecasts. Genuine out-of-sample evaluation of probabilistic demographic forecasts has not been attempted before, to the best of our knowledge.

The aim of this paper is to show how methods for evaluating probabilistic forecasts developed elsewhere can be applied to probabilistic population forecasts. We present and apply scoring functions for prediction intervals, and for simulated samples of future population size and age structures. We illustrate the scoring functions using data for France, the Netherlands, and Norway, and compare probabilistic forecasts computed by various scholars. For the latter two countries, competing probabilistic forecasts are available.

2. Scoring functions

Write the variable for which one computes a forecast as X, with cumulative distribution function (CDF) defined as $F(x) = P(X \leq x)$. Write y for the observed value of X. A scoring function $S(F(x),y)$ assigns a numerical value (a "score") to the forecast $F(x)$, given the observation y. It measures the distance, in some specific way, between the observation and the predictive distribution. Several scoring functions have been proposed (Gneiting and Katzfuss 2014, Gneiting and Raftery 2007). Many of them are based on the following principles: (i) an observation close to the median or the expected value of the predictive distribution gives a good score – the closer the better; (ii) given an observation, a narrow predictive distribution gives a good score – the narrower the better.

There are three main ways to make the results of the probabilistic forecast available: (i) by means of first and second moments, (ii) by means of prediction intervals, and (iii) by means of the sample of simulated results. The type of scoring function depends on these three possibilities. We shall use the following scoring functions.

2.1 Moments-based scoring functions

A scoring function that evaluates the variance of the predictive distribution around the observed value is

$$S(F(x),y) = \mathrm{Var}_y(X) = \sigma^2 + (\mu - y)^2.$$

This scoring function rewards accuracy - when y coincides with μ, the forecast is of optimal quality - and sharpness - a small variance gives a good score. Another advantage is that it is simple to compute.

An alternative scoring function is the Dawid-Sebastiani score

$$DS = \ln(\sigma^2) + (\mu - y)^2/\sigma^2.$$

This scoring function is similar to the variance based score $\mathrm{Var}_y(X)$, but gives somewhat different weight to the forecast variance σ^2. A low variance leads to a good (low) score as long as $\frac{dDS}{d\sigma^2} = \frac{1}{\sigma^2} - \frac{(\mu-y)^2}{\sigma^4} > 0$, or $\sigma > |\mu - y|$.

2.2 Interval scores

Consider a central (1-α).100 % prediction interval$[u,l\,]$, with lower and upper endpoints that are the predictive quantiles at levels $\alpha/2$ and (1- $\alpha/2$), respectively. Then we can define the following score function

$$S_\alpha(l,u;y) = \alpha[(u-l) + (l-y)\mathbb{I}\{y < l\} + (y-u)\mathbb{I}\{y > u\}]\,. \qquad (8)$$

Narrow prediction intervals imply good scores, and the forecaster incurs a penalty, the size of which depends on α, in case the observation misses the interval.[1]

[1] This score function is slightly different from the interval score of Gneiting and Raftery (2007). Their score function is $S_\alpha(l,u;y) = (u-l) + \frac{2}{\alpha}[(l-y)\mathbb{I}\{y < l\} + (y-u)\mathbb{I}\{y > u\}]$. Its drawbacks is that it does not reward, when the observation y is inside the interval $[l,u]$, the forecast with large coverage probability (1 - α) for a certain prediction interval $[l,u]$, compared to a competing forecast with smaller coverage probability for the same interval.

2.3 Scores for distributions

Assume that we have a forecast available in terms of a simulated distribution. Then the CDF is $\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{X_i \leq x\}$, where m is the size of the sample. The Continuous ranked Probability Score CRPS is

$$\text{CRPS}(\hat{F}_m, y) = \frac{2}{m^2}\sum_{i=1}^{m}(X_{(i)} - y)\left(m\mathbb{I}\{y < X_{(i)}\} - i + ½\right),$$

where $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, …, $X_{(m)}$, is the sorted sample; see Gneiting and Raftery (2007) and Jordan et al. (2018).

3. Application

We have evaluated probabilistic population forecasts for several European countries. We focused on total population size and on the population pyramid of the three countries. The data stem from various sources.

1. From the website of "Uncertain Population of Europe" or UPE we used sample paths for the forecasts of the population pyramid for 2010 for the three countries.

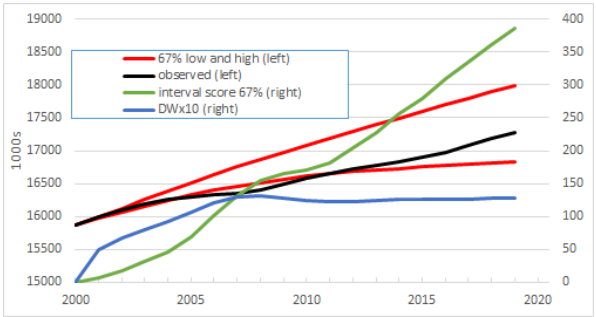
2. Alho and Nikander (2004) report 80 per cent prediction intervals and medians for total population size, amongst others, for each year in the period 2004-2050 for all UPE-countries.

3. For the Netherlands, there is information on prediction intervals for the official probabilistic population forecast with jump-off year 2000; see Statistics Netherlands (2001).

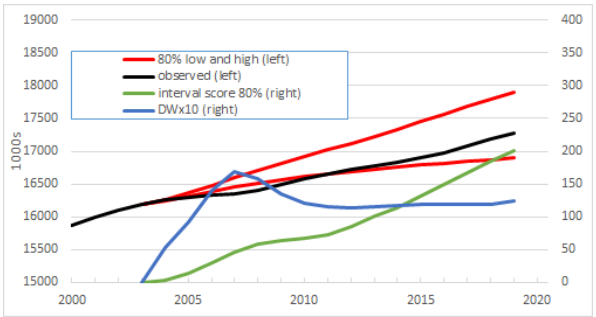
3. For Norway, we use results of the so-called StocProj ("Stochastic Projections") project (Keilman et al. 2002). The jump-off year was 1996. We use interval forecasts for total population size for the years 1997-2019.

Below are some selected results

Figure 1. Scores for total population size, Netherlands. Prediction intervals, observed values (both scale left, in 1000s), interval scores (scale right, in 1000s), and Dawid-Sebastiani (DW) scores.



Statistics Netherlands 1997-2019



UPE 2004-2019

Results for other countries, as well as for age distributions, will be included in the paper.

References

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

Alexopoulos, A., Dellaportas, P., Forster, J.J. (2018) Bayesian forecasting of mortality rates by using latent Gaussian models. Journal of the Royal Statistical Society: Series A (Statistics in Society).

Alho, J. and Nikander, T. (2004) Uncertain population of Europe: Summary results from a stochastic forecast. Available at http://www.stat.fi/tup/euupe/del12.pdf (accessed on 21 March 2019).

Gneiting, T. and Raftery, A. (2007) Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association 102(477) 359-378.

Gneiting, T. and Katzfuss, M. (2014) Probabilistic forecasting. Annual Review of Statistics and Its Applications 1: 125-151.

ISTAT (2018) Il futuro demografico del paese: Previsioni regionali della popolazione residente al 2065 (base 1.1.2017). Report Statistische 3 maggio 2018. Roma: ISTAT.

Jordan A., Krüger F., Lerch S. (2019) Evaluating Probabilistic Forecasts with **scoringRules**. Journal of Statistical Software 90(12).

Keilman, N., Pham, D.Q, Hetland, A. (2002) Why population forecasts should be probabilistic - illustrated by the case of Norway. Demographic Research 6-15 May 2002, 409-454.

Raftery, A., N. Li, H. Ševčíková, P. Gerland, G. Heilig (2012). Bayesian probabilistic population projections for all countries. PNAS 109 (35), 13915-13921.

Shang, H.L., Smith, P., Bijak, J. and Wisniowski, A. (2016) A multilevel functional data method for forecasting population, with an application to the United Kingdom. International Journal of Forecasting 32, 629-649.

Shang, H.L. and Hyndman, R. (2017) Grouped functional time series forecasting: An application to age-specific mortality rates. Journal of Computational and Graphical Statistics 26(2), 330-343.

Statistics New Zealand (2011) National Population Projections: 2011(base)–2061. Bulletin published 19 July 2012, ISSN 1178-0584. Available at http://archive.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalPopulationProjections_HOTP2011.aspx (accessed on 21 March 2019).